

## Mechanisms of action of ELF5 in breast cancer

**Author:**

Piggin, Catherine

**Publication Date:**

2018

**DOI:**

<https://doi.org/10.26190/unsworks/3362>

**License:**

<https://creativecommons.org/licenses/by-nc-nd/3.0/au/>

Link to license to see what you are allowed to do with this resource.

Downloaded from <http://hdl.handle.net/1959.4/59565> in <https://unsworks.unsw.edu.au> on 2024-04-27

# Mechanisms of action of ELF5 in breast cancer

Catherine Louise Piggin

A thesis submitted for the degree of Doctor of Philosophy



Faculty of Medicine

University of New South Wales



Garvan Institute of Medical Research

Cancer Division

January 2018

**THE UNIVERSITY OF NEW SOUTH WALES**  
**Thesis/Dissertation Sheet**

Surname or Family name: Piggin

First name: Catherine

Other name/s: Louise

Abbreviation for degree as given in the University calendar: PhD

School: St Vincent's Clinical School

Faculty: Medicine

Title: Mechanisms of action of ELF5 in breast cancer

**Abstract 350 words maximum:**

The ETS transcription factor ELF5 is a critical regulator of cell fate. In the mammary epithelium, ELF5 drives the development of the ER-negative milk-producing alveolar cells, and the balance between ER and ELF5 transcriptional activity is hypothesised to be a fundamental determinant of cell fate. ELF5-driven transcriptional programs may also function in breast cancer, with previous studies demonstrating roles in basal-like and endocrine-resistant disease. The aim of this thesis was to investigate the transcriptional functions of ELF5 in breast cancer, and the factors that regulate ELF5 activity.

A potential mechanism of ELF5 regulation is through alternative splicing, producing unique protein isoforms. There are four ELF5 isoforms; however little is known about their specific functions. ELF5 expression was comprehensively analysed at the isoform level, using RNA-sequencing data from 6,757 Cancer Genome Atlas samples. In breast cancer, ELF5 alterations were subtype-specific, with the basal subtype demonstrating unique isoform expression changes. Despite differences in protein domains, the *in vitro* functional effects of ELF5 isoforms were similar.

Genome-wide sequencing studies were performed to investigate ELF5 DNA binding sites and transcriptional effects in ER-positive breast cancer cells. ELF5 regulated transcriptional signatures of long-term oestrogen deprivation, suppression of the interferon response, and MYC-regulated gene expression. Increased ELF5 also redistributed the genomic binding sites of the ER pioneer factor FOXA1, representing a novel mechanism by which ELF5 may modulate the oestrogen response.

Finally, interactions with other proteins are essential for specific transcriptional regulation. However, no ELF5-interacting proteins have previously been identified in human breast cancer cells. The protein interactions of chromatin-bound ELF5 were investigated using RIME, identifying DNA-PKcs. A transcriptional model involving ELF5, ER and DNA-PKcs was proposed, with important potential implications for the use of DNA-PKcs inhibitors in breast cancer treatment.

Effective therapeutic targeting of transcription factors depends on a detailed understanding of how transcription factors function, the mechanisms that regulate them, and how these processes are dysregulated in cancer. The new insights into ELF5 function provided by this thesis represent an important contribution towards realising the potential of ELF5 as a therapeutic target in cancer.

**Declaration relating to disposition of project thesis/dissertation**

I hereby grant to the University of New South Wales or its agents the right to archive and to make available my thesis or dissertation in whole or in part in the University libraries in all forms of media, now or here after known, subject to the provisions of the Copyright Act 1968. I retain all property rights, such as patent rights. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

I also authorise University Microfilms to use the 350 word abstract of my thesis in Dissertation Abstracts International (this is applicable to doctoral theses only).

The University recognises that there may be exceptional circumstances requiring restrictions on copying or conditions on use. Requests for restriction for a period of up to 2 years must be made in writing. Requests for a longer period of restriction may be considered in exceptional circumstances and require the approval of the Dean of Graduate Research.

**FOR OFFICE USE ONLY**

Date of completion of requirements for Award:

## **ORIGINALITY STATEMENT**

'I hereby declare that this submission is my own work and to the best of my knowledge it contains no materials previously published or written by another person, or substantial proportions of material which have been accepted for the award of any other degree or diploma at UNSW or any other educational institution, except where due acknowledgement is made in the thesis. Any contribution made to the research by others, with whom I have worked at UNSW or elsewhere, is explicitly acknowledged in the thesis. I also declare that the intellectual content of this thesis is the product of my own work, except to the extent that assistance from others in the project's design and conception or in style, presentation and linguistic expression is acknowledged.'



## **COPYRIGHT STATEMENT**

'I hereby grant the University of New South Wales or its agents the right to archive and to make available my thesis or dissertation in whole or part in the University libraries in all forms of media, now or here after known, subject to the provisions of the Copyright Act 1968. I retain all proprietary rights, such as patent rights. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation. I also authorise University Microfilms to use the 350 word abstract of my thesis in Dissertation Abstract International (this is applicable to doctoral theses only). I have either used no substantial portions of copyright material in my thesis or I have obtained permission to use copyright material; where permission has not been granted I have applied/will apply for a partial restriction of the digital copy of my thesis or dissertation.'

## **AUTHENTICITY STATEMENT**

'I certify that the Library deposit digital copy is a direct equivalent of the final officially approved version of my thesis. No emendation of content has occurred and if there are any minor variations in formatting, they are the result of the conversion to digital format.'

## Abstract

The ETS transcription factor ELF5 is a critical regulator of cell fate. In the mammary epithelium, ELF5 drives the development of the ER-negative milk-producing alveolar cells, and the balance between ER and ELF5 transcriptional activity is hypothesised to be a fundamental determinant of cell fate. ELF5-driven transcriptional programs may also function in breast cancer, with previous studies demonstrating roles in basal-like and endocrine-resistant disease. The aim of this thesis was to investigate the transcriptional functions of ELF5 in breast cancer, and the factors that regulate ELF5 activity.

A potential mechanism of ELF5 regulation is through alternative splicing, producing unique protein isoforms. There are four ELF5 isoforms; however little is known about their specific functions. ELF5 expression was comprehensively analysed at the isoform level, using RNA-sequencing data from 6,757 Cancer Genome Atlas samples. In breast cancer, ELF5 alterations were subtype-specific, with the basal subtype demonstrating unique isoform expression changes. Despite differences in protein domains, the *in vitro* functional effects of ELF5 isoforms were similar.

Genome-wide sequencing studies were performed to investigate ELF5 DNA binding sites and transcriptional effects in ER-positive breast cancer cells. ELF5 regulated transcriptional signatures of long-term oestrogen deprivation, suppression of the interferon response, and MYC-regulated gene expression. Increased ELF5 also redistributed the genomic binding sites of the ER pioneer factor FOXA1, representing a novel mechanism by which ELF5 may modulate the oestrogen response.

Finally, interactions with other proteins are essential for specific transcriptional regulation. However, no ELF5-interacting proteins have previously been identified in human breast cancer cells. The protein interactions of chromatin-bound ELF5 were investigated using RIME, identifying DNA-PKcs. A transcriptional model involving ELF5, ER and DNA-PKcs was proposed, with important potential implications for the use of DNA-PKcs inhibitors in breast cancer treatment.

Effective therapeutic targeting of transcription factors depends on a detailed understanding of how transcription factors function, the mechanisms that regulate them, and how these processes are dysregulated in cancer. The new insights into ELF5 function provided by this thesis represent an important contribution towards realising the potential of ELF5 as a therapeutic target in cancer.

## **Acknowledgements**

Thank you to my supervisor Professor Chris Ormandy for giving me the best start in science I could have asked for. I am grateful to all those who have helped me along the way, including my co-supervisor Professor Susan Clark, Dr Heather Lee, Dr David Gallego-Ortega, Dr Samantha Oakes, Dr Daniel Roden, Anita Ledger, Marcelo Sergio, Gillian Lehrbach, Andrew Law, Hayley Cullen, Dr Adelaide Young and Stephanie Allerdice (my PhD partners in crime!), and other members of the Cancer Biology group past and present.

Special thanks go to those who have provide valuable assistance with the studies described in this thesis, including Dr Daniel Roden (bioinformatics), Dr Jason Carroll (ChIP-seq and RIME), and those at The Australian Proteome Analysis Facility and the Ramaciotti Centre for Genomics.

Finally, thank you to my parents for their wonderful support and many cups of coffee.

## List of Abbreviations and Symbols

3'	3 prime (denoting the downstream end of the DNA molecule with a hydroxyl group on carbon 3)
3C	Chromosome conformation capture
5'	5 prime (denoting the upstream end of the DNA molecule with a phosphate on carbon 5)
5mC	5-methylcytosine
A, T, G, C	Adenine, thymine, guanine, cytosine
ABP	AU-rich element binding protein
AI	Aromatase inhibitor
AIRE	Autoimmune regulator
AKT	AKT serine/threonine kinase
AMBIC	Ammonium hydrogen carbonate
AMPK	AMP-activated protein kinase
ANOVA	Analysis of variance
AP1	Activator protein 1
APA	Alternative polyadenylation
AR	Androgen receptor
ARE	AU-rich element
ATM	Ataxia telangiectasia mutated serine/threonine kinase
ATP	Adenosine triphosphate
ATR	Ataxia telangiectasia and RAD3-related serine/threonine kinase
bp	Base pair(s)
BPTF	Bromodomain PHD finger transcription factor
BRCA1, 2	Breast cancer susceptibility type 1 and type 2 proteins
BRD3, 4	Bromodomain-containing protein 3, 4
CaMKII	Calcium / calmodulin-dependent kinase II
CBX1, 5	Chromobox 1, 5
CDK9	Cyclin-dependent kinase 9
CEAS	<i>Cis</i> -regulatory Element Annotation System
CEBPa, b	CCAAT/enhancer binding protein alpha, beta
CGI(s)	CpG island(s)
ChEA	ChIP enrichment analysis
CHD3, 4	Chromodomain helicase DNA binding protein 3, 4
ChIP-seq	Chromatin immunoprecipitation sequencing
CHK1, 2	Serine/threonine-protein kinase CHK1, 2
CK1,2	Casein kinase 1,2
CNRQ	Calibrated normalised relative quantity
Co-IP	Co-immunoprecipitation

CpG	Cytosine-phosphate-guanine (indicating linear CG dinucleotide on a single DNA strand)
CREB1	cAMP response element binding protein 1
CREBBP	CREB binding protein (also known as CBP)
CRISPR	Clustered regularly interspaced short palindromic repeat
CSK21	Casein kinase 2 subunit alpha
CTCF	CCCTC-binding factor
DBD	DNA-binding domain
DDR	DNA-damage response
DEG	Differentially expressed gene
DNA	Deoxyribonucleic acid
DNA-PK	DNA-dependent protein kinase (composed of Ku70, Ku80 and catalytic sub-units)
DNA-PKcs	DNA-dependent protein kinase catalytic sub-unit
DNMT1, 3A, 3B, 3L	DNA methyltransferase 1, 3A, 3B, 3L
Dox	Doxycycline
DPE	Downstream promoter element
DSB	Doubled-stranded break (of DNA)
DSIF	DRB sensitivity inducing factor
E2	Oestradiol
EGF(R)	Epidermal growth factor (receptor)
EHF	ETS homologous factor
ELAVL1	ELAV-like binding protein 1
ELF1, 2, 3, 4, 5	E74 like factor transcription factor 1, 2, 3, 4, 5
ELK1, 3, 4	ELK1, 3, 4, ETS transcription factor
EMT	Epithelial to mesenchymal transition
ENCODE	Encyclopedia of DNA elements
ER	Oestrogen receptor alpha
ERBB2	erb-b2 receptor tyrosine kinase 2 (also known as HER2)
ERF	ETS2 repressor factor
ERG	ERG transcription factor
ESC	Embryonic stem cell
ETS	E26 transforming sequence
ETS1	ETS proto-oncogene 1
ETS2	ETS proto-oncogene 2
ETV1, 2, 3, 4, 5, 6, 7	ETS variant 1, 2, 3, 4, 5, 6 (TEL), 7
ETV3L	ETS variant 3 like
FC	Fold change
FDR	False discovery rate
FEV	FEV, ETS transcription factor

FLI1	FLI-1 proto-oncogene
FOS	Fos proto-oncogene, AP-1 transcription factor subunit
FOXA1, A2, C1, M1, O3, O4, P1	Forkhead box A1, A2, C1, M1, O4, P1
FPKM	Fragments Per Kilobase of transcript per Million mapped reads
GABPA	GA-binding protein transcription factor alpha subunit
GATA1, 3	GATA binding protein 1, 3
GBS	Glucocorticoid receptor binding sequence
GO (BP, CC)	Gene ontology (biological process, cellular component)
GR	Glucocorticoid receptor
GRE	Glucocorticoid response element
GREAT	Genomic Regions Enrichment of Annotations Tool
GRHL2	Grainyhead-like transcription factor 2
GSEA	Gene set enrichment analysis
H3K4me1/2/3	Histone H3 lysine 4 mono-/di-/tri-methylation
H3K9me1	Histone 3 lysine 9 monomethylation
H3K27ac	Histone H3 lysine 27 acetylation
H3K27me3	Histone H3 lysine 27 trimethylation
HAT	Histone acetyltransferase
HDAC	Histone deacetylase
HER2	erb-b2 receptor tyrosine kinase 2 (also known as ERBB2)
hESC(s)	Human embryonic stem cell(s)
HR	Homologous recombination
HSP90a	Heat shock protein 90a
IFITM1	Interferon induced transmembrane protein 1
IFN	Interferon
IP	Immunoprecipitation
IRF17, 9	Interferon regulatory factor 1,7, 9
IGF1, 2	Insulin-like growth factor 1, 2
IGF1R	Insulin-like growth factor 1 receptor
IGFBP5	Insulin like growth factor binding protein 5
IGV	Integrative Genomics Viewer
Inr	Initiator element
iPSC	Induced pluripotent stem cell
JUN	Jun proto-oncogene, AP-1 transcription factor subunit
KDM1A	Lysine demethylase 1A (also known as LSD1)
kb	Kilobase(s)
KDa	Kilodalton(s)
KLF4	Kruppel like factor 4
LBD	Ligand-binding domain

LTED	Long-term oestrogen deprivation
MAPK	Mitogen-activated protein kinase
MAX	MYC-associated factor X
MDSC	Myeloid-derived suppressor cell
mESC(s)	Mouse embryonic stem cell(s)
MET	Mesenchymal to epithelial transition
miRNA	MicroRNA
MLL3	Mixed-lineage leukaemia 3
mRNA	Messenger RNA
MS	Mass spectrometry
MSigDB	Molecular signatures database
mTOR	Mammalian target of rapamycin
MYC	MYC proto-oncogene, bHLH transcription factor
NCOR2	Nuclear receptor co-repressor 2 (also known as SMRT)
NEDD8	Neural precursor cell expressed developmentally downregulated protein 8
NELF	Negative elongation factor
NES	Nuclear export sequence
NF $\kappa$ B	Nuclear factor kappa B
NHEJ	Non-homologous end joining
NPC(s)	Nuclear pore complex(es)
NURD	Nucleosome remodeling and histone deacetylation complex
NURF	Nucleosome-remodelling factor complex
OCT4	POU class 5 homeobox 1
OMIM	Online Mendelian Inheritance in Man
p53	Tumour protein p53
PAM50	Prediction Analysis of Microarray 50-gene classifier
PARP1	Poly(ADP-ribose) polymerase 1
PAX3, 5	Paired box 3, 5
PDGF	Platelet-derived growth factor
PCR	Polymerase chain reaction
PDX1	Pancreatic and duodenal homeobox 1
PIC	Pre-initiation complex
PIP	Prolactin-induced protein
PLA	Proximity ligation assay
PLK1	Polo-like kinase 1
PNT	Pointed domain
Pol II	RNA Polymerase II
Pol II CTD	RNA Polymerase II C-terminal domain
POU5F1	POU class 5 homeobox 1 (also known as OCT4)

PP6	Protein phosphatase 6
PR	Progesterone receptor
P-TEFb	Positive transcription elongation factor b
PTM(s)	Post-translational modification(s)
PU.1	Spi1 proto-oncogene (also known as SPI1)
PyMT	Polyoma middle T
qPCR	Quantitative polymerase chain reaction
RAN	RAN, member RAS oncogene family
RB1	RB transcriptional corepressor 1
RIME	Rapid immunoprecipitation of endogenous protein
RNA	Ribonucleic acid
rRNA	Ribosomal RNA
RUNX1	Runt-related transcription factor 1
SAM	Sterile Alpha Motif domain
SCID	Severe combined immunodeficiency
SEC16A	Protein transport protein Sec16A
SIN3A	Sin3 transcription regulator family member A
siRNA	Small interfering ribonucleic acid
SMARCA2	SWI/SNF related matrix associated actin dependent regulator of chromatin, subfamily a, member 2 (also known as Brahma homolog, BRM)
SMARCA4	SWI/SNF related matrix associated actin dependent regulator of chromatin, subfamily a, member 4 (also known as BRG1)
SMAD2	SMAD family member 2
SNAI1, 2	Snail family transcriptional repressor 1, 2
SOX2	SRY-box 2
SP1	Sp1 transcription factor
SPDEF	SAM pointed domain containing ETS transcription factor
SPI1	Spi1 proto-oncogene (also known as PU.1)
SPIB	SPI-B transcription factor
SPIC	SPI-C transcription factor
SRC1, 2, 3	Steroid receptor co-activator 1, 2, 3
SRCAP	Snf2 related CREBBP activator protein
STAT1, 5	Signal transducer and activator of transcription 1, 5
SUV39H1	Suppressor of variegation 3-9 homolog 1
SUZ12	SUZ12 polycomb repressive complex 2 subunit
SV40	Simian virus 40
SWI/SNF	SWItch / Sucrose Non Fermentable
TAF	TATA-binding protein associated factor
TAMR	Tamoxifen-resistant



TBP	TATA-binding protein
TCGA	The Cancer Genome Atlas
Tfap2C	Transcription factor AP-2, gamma
TGF-b	Transforming growth factor-beta
TLE1	Transducin like enhancer of split 1
TLE3	Transducin like enhancer of split 3 (also known as GRG3)
TOP2A, B	DNA topoisomerase 2-alpha, beta
TPM	Transcripts Per Million, a proportional measure of abundance correcting for transcript length
TR	Thyroid hormone receptor
tRNA	Transfer RNA
TSC(s)	Trophoblast stem cell(s)
TSS	Transcription start site
UTR	Untranslated region
VEGF	Vascular Endothelial Growth Factor
<i>v-ets</i>	Viral E26 transofrmation-specific sequence
VTCN1	V-set domain containing T cell activation inhibitor 1
XRCC5, 6	X-ray repair cross-complementing proteins 5, 6 (also known as Ku70 and Ku80/Ku86)
YWHAB, YWHAG	Tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation proteins beta, gamma
ZEB1, 2	Zinc finger E-box binding homeobox 1, 2
ZNF	Zinc finger protein

# List of Figures and Tables

## Chapter 1

Figure 1.1: The general process of transcription consists of several stages .....	8
Figure 1.2: Multiple mechanisms of action of eukaryotic transcription factors.....	12
Figure 1.3: Cell-type-specific enhancers have characteristic chromatin features.....	15
Figure 1.4: Super-enhancers can be defined based on Mediator binding signal.....	17
Figure 1.5: Mechanisms of transcription factor co-operativity may be direct or indirect .....	21
Figure 1.6: Model of GATA1 and PU.1 cross-antagonism during haematopoietic development .....	27
Figure 1.7: Co-factors as readers and effectors of context-dependent changes in glucocorticoid receptor conformation .....	29
Table 1.1: Examples of common post-translational histone modifications .....	41
Figure 1.8: Pioneer factors interact with nucleosomal DNA and “bookmark” enhancers for future activation .....	44
Figure 1.9: Multiple signalling cascades are initiated on binding of prolactin to the long isoform of the prolactin receptor.....	53
Table 1.2: The nuclear receptor superfamily.....	56
Table 1.3: Post-translational modifications (PTMs) of transcription factors .....	57
Figure 1.10: Transcript variants are produced by alternative promoters, alternative polyadenylation sites and alternative splicing .....	65
Figure 1.11: The human <i>TP53</i> gene encodes 9 mRNA transcripts, producing 12 protein isoforms with unique domain structures .....	67
Figure 1.12: Common microRNA-transcription factor regulatory loops .....	75
Figure 1.13: Interactions between transcription factors and methylated DNA.....	85
Table 1.4: The intrinsic subtypes of breast cancer.....	95
Figure 1.14: Two models of intertumoural heterogeneity.....	96
Figure 1.15: The cellular organisation of the normal mammary gland .....	97
Figure 1.16: Model of the human mammary epithelial cell hierarchy linked to cancer subtype .....	99
Table 1.5: The ETS transcription factor family .....	103
Figure 1.17: Molecular mechanisms of ETS factor activation and inactivation in cancer .....	105
Table 1.6: Molecular mechanisms of ETS factor dysregulation in cancer .....	107
Figure 1.18: The DNA-damage response .....	116
Figure 1.19: Functional domains and structure of DNA-PKcs .....	118
Table 1.7: Emerging roles for DNA-PKcs.....	121
Table 1.8: Examples of transcription factors interacting with DNA-PKcs.....	122
Figure 1.20: The Role of the DNA-Damage Response in Cancer .....	128
Table 1.9: DNA-PKcs expression and activity in clinical cancer studies.....	133

## Chapter 2

Table 2.1: Cell lines .....	139
Table 2.2: Western blot experiments .....	147
Table 2.3: Western blot antibodies .....	149
Table 2.4: Quantitative PCR experiments .....	152
Table 2.5: Roche qPCR assays (related to Figure 3.17) .....	154
Table 2.6: Taqman qPCR assays (related to Figure 3.17) .....	158
Table 2.7: Quantitative PCR assays (related to Figure 5.18) .....	160

## Chapter 3

Figure 3.1: ELF5 isoforms are produced by alternative promoter usage and splicing .....	168
Figure 3.2: <i>ELF5</i> annotations .....	169
Table 3.1: Summary of all TCGA RNA-sequencing samples used in analysis .....	171
Figure 3.3: <i>ELF5</i> isoforms are differentially expressed in normal tissues (quantile normalised counts) .....	172
Table 3.2: <i>ELF5</i> splice variant proportions in normal tissues (based on TPM values) .....	173
Figure 3.4: <i>ELF5</i> expression is significantly altered in cancer .....	174
Figure 3.5: Additional RNA-Seq datasets for normal tissues .....	175
Figure 3.6: <i>ELF5</i> expression is altered in breast cancer in a subtype-specific manner .....	177
Figure 3.7: Stage compared to <i>ELF5</i> expression in breast cancer subtypes .....	179
Figure 3.8: <i>ELF5</i> isoform expression in normal tissues and cancer (TPM values) .....	180
Figure 3.9: ETS family gene expression in normal breast and breast cancer (TPM) .....	181
Figure 3.10: Expression of other ETS family members is also altered in breast cancer, with the basal subtype having a distinct ETS expression profile .....	183
Figure 3.11: ETS family expression gene expression in normal breast and breast cancer subtypes (normalised counts) .....	184
Figure 3.12: <i>ELF5</i> mRNA and protein expression in breast cancer cell lines .....	186
Figure 3.13: Alterations in cell line <i>ELF5</i> isoform levels result in a similar phenotype, characterised by decreased cell number and decreased expression of oestrogen- related proteins .....	188
Figure 3.14: Phenotype of pHUSH- <i>ELF5</i> -V5 breast cancer clonal cell lines .....	189
Figure 3.15: <i>ELF5</i> Isoform 3 expression does not alter Isoform 2 subcellular localisation .....	190
Figure 3.16: qPCR workflow .....	191
Figure 3.17: <i>ELF5</i> isoforms have a similar transcriptional effect in T47D and MDA-MB-231 cell lines .....	193
Table 3.3: Clonal cell lines used in qPCR panel .....	195
Additional Figure 3.1: <i>ELF5</i> expression is significantly altered in cancer - results from edgeR differential expression analysis .....	199
Additional Figure 3.2: Expression of other ETS family members is also altered in breast cancer, with the basal subtype having a distinct ETS expression profile - results from edgeR differential expression analysis .....	200

## Chapter 4

Figure 4.1: Model of the mammary epithelial hierarchy .....	219
Figure 4.2: Visualisation of the transcriptional functions of ELF5 in MCF7-ELF5-V5 cells discovered using RNA sequencing .....	223
Figure 4.3: Gene ontology analysis of differentially expressed genes .....	226
Table 4.1: Numbers of Differentially Expressed Genes in MCF7-ELF5-V5 RNA-seq and Microarray Experiments .....	227
Figure 4.4: Comparison of the MCF7-ELF5-V5 RNA-seq and microarray experiments .....	228
Figure 4.5: Comparison of the ELF5 transcriptional networks discovered using RNA sequencing and microarray .....	230
Figure 4.6: Selected gene sets from the Cytoscape network .....	232
Figure 4.7: Hallmark collection gene set enrichment analysis .....	237
Figure 4.8: Oncogenic signatures gene set enrichment analysis .....	240
Figure 4.9: “Enrichr” analysis of RNA-seq-identified differentially expressed genes .....	242
Figure 4.10: ELF5 ChIP-seq summary .....	245
Figure 4.11: Functional analysis of ELF5 ChIP-seq peaks .....	246
Figure 4.12: Integration of ELF5 RNA-seq and ChIP-seq to identify direct ELF5 target genes .....	251
Figure 4.13: FOXA1-ELF5-low ChIP-seq summary .....	255
Figure 4.14: FOXA1-ELF5-high ChIP-seq summary .....	256
Figure 4.15: Functional analysis of FOXA1-ELF5-low and FOXA1-ELF5-high ChIP-seq peaks .....	258
Figure 4.16: ELF5-driven redistribution of FOXA1 binding .....	261
Figure 4.17: ChIP-seq signals at genomic locations of redistributed FOXA1 binding .....	263
Figure 4.18: Examples of weak (background) and strong ChIP-seq signals .....	265
Figure 4.19: Functional analysis of gained FOXA1 binding sites .....	267
Figure 4.20: Analysis of ELF5 co-binding at gained FOXA1 binding sites .....	269
Figure 4.21: Analysis of lost FOXA1 binding sites .....	271
Figure 4.22: Potential mechanisms of ELF5-driven redistribution of FOXA1 binding .....	290
Additional Figure 4.1: Enlarged Cytoscape sub-networks (based on Figure 4.1) .....	293
Additional Table 4.1: Differentially expressed genes (up-regulated) in MCF7-ELF5-V5 RNA-seq .....	300
Additional Table 4.2: Differentially expressed genes (down-regulated) in MCF7-ELF5-V5 RNA-seq .....	306
Additional Tables 4.3: Heatmap gene symbols and log2 fold change values .....	313

## Chapter 5

Figure 5.1: Rapid Immunoprecipitation of Endogenous Protein (RIME) purifies cross- linked transcriptional complexes .....	346
Table 5.1: Numbers of proteins identified in ELF5-V5 and IgG control RIME replicates .....	349
Table 5.2: ELF5-interacting proteins identified by RIME .....	351

Figure 5.2: ELF5-V5 RIME purifies multiple proteins in MCF7 luminal breast cancer cells.....	354
Figure 5.3: ELF5-V5 RIME proteins overlap with published mouse Elf5 and human ER interacting proteins .....	356
Table 5.3: Common proteins identified in ELF5-V5 RIME (MCF7 cells) and Elf5 MS (mouse trophoblastic stem cells) .....	357
Table 5.4: Common proteins identified in ELF5-V5 and ER RIME (MCF7 cells).....	358
Figure 5.4: DNA-PKcs is significantly altered in cancer .....	361
Figure 5.5: ELF5-V5 and DNA-PKcs co-immunoprecipitate in MCF7 cells .....	364
Figure 5.6: The immunofluorescence-based Proximity Ligation Assay (PLA) identifies interacting proteins .....	365
Figure 5.7: ELF5-V5 single-antibody PLA optimisation .....	366
Figure 5.8A: DNA-PKcs CST single-antibody PLA optimisation.....	367
Figure 5.8B: DNA-PKcs Thermo single-antibody PLA optimisation.....	368
Figure 5.9A: Proximity ligation assays corroborate the ELF5-DNA-PKcs nuclear interaction (V5 + DNA-PKcs CST combination).....	370
Figure 5.9B: Proximity ligation assays corroborate the ELF5-DNA-PKcs nuclear interaction (V5 + DNA-PKcs Thermo Fisher combination) .....	371
Figure 5.10: Quantification of ELF5-DNA-PKcs PLA signals in MCF7 cell lines .....	373
Figure 5.11: ELF5 is a phosphoprotein <i>in vivo</i> .....	376
Figure 5.12: Efficient knockdown of DNA-PKcs can be achieved in breast cancer cells.....	378
Figure 5.13: DNA-PKcs knockdown does not alter ELF5 phosphorylation.....	380
Figure 5.14: DNA-PKcs knockdown significantly reduces cell number in luminal breast cancer cell lines .....	382
Figure 5.15A: DNA-PKcs knockdown alters cell phenotype over a 4-day timecourse (MCF7 cells) .....	383
Figure 5.15B: DNA-PKcs knockdown alters cell phenotype over a 4-day timecourse (T47D cells) .....	384
Figure 5.15C: DNA-PKcs knockdown alters cell phenotype over a 4-day timecourse (MDA-MB-231 cells) .....	385
Figure 5.16A: ELF5 overexpression does not affect the DNA-PKcs knockdown phenotype (MCF7 cells) .....	387
Figure 5.16B: ELF5 overexpression does not affect the DNA-PKcs knockdown phenotype (T47D cells) .....	388
Figure 5.16A: ELF5 overexpression does not affect the DNA-PKcs knockdown phenotype (MDA-MB-231 cells) .....	389
Figure 5.17A: Enlarged images of DNA-PKcs knockdown phenotype in luminal breast cancer cell lines (MCF7 cells) .....	390
Figure 5.17B: Enlarged images of DNA-PKcs knockdown phenotype in luminal breast cancer cell lines (T47D cells) .....	391
Figure 5.18: Knockdown of <i>DNA-PKcs</i> affects expression of ELF5-regulated genes .....	398

Figure 5.19A: DNA-PKcs knockdown affects expression of ELF5 and other breast cancer-associated proteins (MCF7 cells) .....	406
Figure 5.19B: DNA-PKcs knockdown affects expression of ELF5 and other breast cancer-associated proteins (T47D cells) .....	407
Figure 5.19C: DNA-PKcs knockdown affects expression of ELF5 and other breast cancer-associated proteins (MDA-MB-231 cells) .....	408
Figure 5.20: ELF5 decreases DNA-PKcs expression in clonal cell lines .....	409
Figure 5.21: ELF5 does not alter DNA-PKcs phosphorylation in pooled cell lines .....	410
Figure 5.22: E2-induced ER phosphorylation is not affected by ELF5 overexpression .....	412
Figure 5.23: DNA-PKcs expression may have subtype-specific effects on breast cancer survival .....	417
Figure 5.24: Model of ELF5, ER and DNA-PKcs interaction in breast cancer .....	428

## Publications Arising From Thesis

Gallego-Ortega, D., Oakes, S. R., Lee, H. J., Piggin, C. L. and Ormandy, C. J. (2013). Elf5, normal mammary development and the heterogeneous phenotypes of breast cancer. *Breast Cancer Management*, 2(6), 489-498. doi: 10.2217/bmt.13.50

Kalyuga, M., Gallego-Ortega, D., Lee, H. J., Roden, D. L., Cowley, M. J., Caldon, C. E., *et al.* (2012). ELF5 suppresses estrogen sensitivity and underpins the acquisition of antiestrogen resistance in luminal breast cancer. *PLoS Biol*, 10(12), e1001461. doi: 10.1371/journal.pbio.1001461

Law, A. M. K., Yin, J. X. M., Castillo, L., Young, A. I. J., Piggin, C., Rogers, S., *et al.* (2017). Andy's Algorithms: new automated digital image analysis pipelines for FIJI. *Sci Rep*, 7(1), 15717. doi: 10.1038/s41598-017-15885-6

Piggin, C. L., Roden, D. L., Gallego-Ortega, D., Lee, H. J., Oakes, S. R. and Ormandy, C. J. (2016). ELF5 isoform expression is tissue-specific and significantly altered in cancer. *Breast Cancer Res*, 18(1), 4. doi: 10.1186/s13058-015-0666-0

# Table of Contents

<b>Abstract</b> .....	<b>i</b>
<b>Acknowledgements</b> .....	<b>ii</b>
<b>List of Abbreviations and Symbols</b> .....	<b>iii</b>
<b>List of Figures and Tables</b> .....	<b>ix</b>
<b>Publications Arising From Thesis</b> .....	<b>xiv</b>
<b>Chapter 1: Introduction and Background</b> .....	<b>1</b>
<b>Introduction</b> .....	<b>1</b>
Transcription Factors .....	1
Targeting Transcription Factors Therapeutically .....	2
ELF5 and Breast Cancer .....	4
Thesis overview .....	5
<b>1.1 How Transcription Factors Function</b> .....	<b>7</b>
General Process of Transcription .....	7
Interactions with the Core Transcriptional Machinery .....	10
Enhancers .....	13
Super-Enhancers .....	16
DNA-Binding Specificity .....	18
Transcription Factor Co-operativity .....	20
Transcription Factor Competition .....	25
Interactions with Co-factors .....	28
Chromatin Remodelling .....	33
Histone Modifications .....	36
Pioneer Functions .....	42
DNA Methylation .....	48
<b>1.2 Regulation of Transcription Factors</b> .....	<b>52</b>
Signalling Pathways .....	52
Ligand Binding .....	54
Post-Translational Modifications .....	56
DNA Binding Sites .....	60
Protein-Protein Interactions .....	62
Protein Levels .....	62
Transcript Variants .....	64
Auto-inhibition .....	71
Subcellular Localisation .....	72
Non-Coding RNA .....	73
Chromatin Structure .....	77



Histone Modifications .....	79
Histone Exchange and Variants.....	82
DNA Methylation .....	84
<b>1.3 Breast Cancer .....</b>	<b>90</b>
Two Clinical Challenges: Endocrine-resistant Disease and “Triple-Negative”	
Breast Cancers .....	90
Clinical and Molecular Subtypes .....	91
Mammary Development and Breast Cancer Subtypes.....	96
<b>1.4 Transcription Factors in Cancer.....</b>	<b>101</b>
The ETS Transcription Factor Family .....	101
ETS Factor Specificity.....	104
ETS Factors in Cancer .....	104
The ETS Transcription Factor ELF5 .....	109
ELF5 Regulates Cell Fate .....	109
ELF5 in Cancer .....	110
<b>1.5 DNA Repair Proteins: A Novel Class of Transcriptional Regulators .....</b>	<b>113</b>
DNA Repair and Transcription .....	113
The DNA-Damage Response .....	115
The DNA-Dependent Protein Kinase Catalytic Sub-Unit .....	117
The Role of DNA-PKcs in DNA Repair .....	119
Emerging Roles for DNA-PKcs .....	120
DNA-Damage Response Proteins in Cancer .....	125
DNA-PKcs in Cancer.....	129
DNA-PKcs in Breast Cancer .....	136
<b>Chapter 2: Materials and Methods.....</b>	<b>138</b>
<b>Cell-Based Methods .....</b>	<b>138</b>
Stable cell line generation .....	138
Cell lines and treatments.....	138
Cell number assessment.....	140
Transient retroviral infection.....	140
Immunofluorescence.....	140
siRNA transfection .....	141
Treatment with ionising radiation .....	141
Proximity ligation assays (PLAs).....	141
<b>Image-Based Methods .....</b>	<b>142</b>
PLA image acquisition.....	142
PLA image analysis.....	142
<b>Protein-Based Methods .....</b>	<b>143</b>
Rapid Immunoprecipitation of Endogenous Protein (RIME) .....	143
Co-immunoprecipitations for western blot.....	145

Phosphoprotein purification .....	145
Western blots .....	145
<b>RNA-Based Methods .....</b>	<b>150</b>
End-point PCR .....	150
Quantitative PCR .....	150
RNA-sequencing (MCF7-ELF5-V5 cells) .....	161
<b>DNA-Based Methods .....</b>	<b>161</b>
ChIP-seq .....	161
<b>Computational Methods .....</b>	<b>162</b>
TCGA RNA-seq analysis .....	162
Additional sequencing datasets analysis .....	163
MCF7-ELF5-V5 RNA-sequencing analysis .....	163
FOXA1 and ELF5 ChIP-seq analysis .....	164
Association of ChIP-seq peaks with direct target genes .....	165
Functional analyses of gene and protein lists .....	165
Gene ID conversions .....	165
Correlation analysis .....	166
Graphics .....	166
Phosphosite prediction .....	166
Survival analysis .....	166
<b>Chapter 3: ELF5 Isoforms .....</b>	<b>167</b>
<b>Introduction .....</b>	<b>167</b>
<b>Results .....</b>	<b>168</b>
<i>ELF5</i> isoforms are differentially expressed in normal tissues .....	168
<i>ELF5</i> expression is significantly altered in cancer .....	173
<i>ELF5</i> expression is altered in breast cancer in a subtype-specific manner .....	176
Expression of other ETS family members is also altered in breast cancer, with the basal subtype having a distinct ETS expression profile .....	178
Alterations in cell line <i>ELF5</i> isoform levels result in a similar phenotype, characterised by decreased cell number, decreased oestrogen-related proteins and nuclear localisation .....	186
<i>ELF5</i> isoforms have a similar transcriptional effect in T47D and MDA-MB-231 cell lines .....	190
<b>Discussion .....</b>	<b>196</b>
<b>Additional Figures .....</b>	<b>199</b>
<b>Appendix: ELF5 isoform expression is tissue-specific and significantly altered     in cancer (Piggin <i>et al</i>, 2016) .....</b>	<b>201</b>

<b>Chapter 4: Genome-wide Studies of ELF5 Action .....</b>	<b>219</b>
<b>Introduction .....</b>	<b>219</b>
ELF5 regulates cell fate in normal development and breast cancer .....	219
<b>Results .....</b>	<b>222</b>
Identification of ELF5-regulated genes using RNA sequencing .....	222
Functional signatures of ELF5 overexpression .....	222
Comparison with MCF7-ELF5-V5 microarray .....	226
Hallmark gene set enrichment analysis .....	234
Oncogenic signatures gene set enrichment analysis .....	237
Enrichr analysis of ELF5-regulated genes .....	240
Identification of ELF5 genomic binding sites in MCF7-ELF5-V5 cells using ChIP-seq .....	243
GREAT functional analysis of ELF5 ChIP-seq consensus peaks .....	245
Enrichr functional analysis of ELF5 ChIP-seq peaks .....	248
Identification of the direct regulatory targets of ELF5 .....	250
FOXA1 genomic binding sites in the context of low and high ELF5 expression .....	253
Functional analysis of FOXA1- ELF5-low and ELF5-high ChIP-seq consensus peaks .....	254
Identification of ELF5-induced changes in FOXA1 binding sites .....	259
Functional analysis of repartitioned FOXA1 binding sites .....	266
<b>Discussion .....</b>	<b>272</b>
Overview .....	272
Transcriptional signatures of long-term oestrogen deprivation .....	272
The interferon response .....	277
ELF5 and MYC .....	280
Single-cell heterogeneity and the dynamics of differentiation .....	285
The relationship between ELF5 and FOXA1 .....	286
FOXA1 is known to regulate ER-chromatin interactions .....	287
ELF5 alters FOXA1 genomic binding .....	287
Potential mechanisms of ELF5-driven redistribution of FOXA1 binding .....	289
<b>Additional Figures and Tables .....</b>	<b>293</b>
<b>Chapter 5: The Interaction Between ELF5 and DNA-PKcs .....</b>	<b>345</b>
<b>Introduction .....</b>	<b>345</b>
<b>Results .....</b>	<b>349</b>
Purification of ELF5-V5-associated proteins using RIME .....	349
Comparison of ELF5 and ER interactomes .....	355
DNA-PKcs expression and alterations in breast cancer .....	359
Validation of the interaction between ELF5 and DNA-PKcs using co-immunoprecipitation .....	363

Validation of the interaction between ELF5 and DNA-PKcs using Proximity	
Ligation Assays .....	364
Phosphorylation of ELF5.....	374
Optimisation of siRNA-mediated knockdown of DNA-PKcs in breast cancer	
cell lines .....	376
ELF5-V5 phosphorylation in DNA-PKcs-knockdown cells.....	378
Phenotype of DNA-PKcs-knockdown cells .....	381
Effects of DNA-PKcs knockdown on ELF5 transcriptional function	
(gene expression) .....	392
Effects of DNA-PKcs knockdown on ELF5 transcriptional function	
(protein expression) .....	405
ELF5 regulation of DNA-PKcs .....	408
Interplay between ELF5, DNA-PKcs and ER.....	411
DNA-PKcs expression and breast cancer survival .....	413
<b>Discussion .....</b>	<b>419</b>
Overview .....	419
Advantages and limitations of RIME .....	419
ELF5 is a phosphoprotein <i>in vivo</i> .....	421
DNA-PKcs as a modulator of ELF5 transcriptional activity.....	422
Site of interaction between ELF5 and DNA-PKcs.....	425
Mutual regulation of DNA-PKcs and ELF5.....	426
A model for DNA-PKcs, ELF5 and ER transcriptional regulation .....	427
Challenges in interpreting survival data .....	428
Therapeutic implications of the interaction between DNA-PKcs and ELF5.....	429
Unanswered questions about DNA-PKcs in transcriptional regulation .....	431
<b>Chapter 6: Conclusions and Future Directions .....</b>	<b>433</b>
<b>References .....</b>	<b>438</b>

# Chapter 1: Introduction and Background

## Introduction

### Transcription Factors

Transcription factors are the integrators of multiple signalling pathways, converting internal and external stimuli into changes in gene expression. This role begins at the earliest stage of development, with transcription factors guiding the differentiation of the blastocyst into the embryonic and placental cell lineages. Ultimately, every one of the 37 trillion cells in the adult human body (Bianconi *et al.*, 2013) is defined by the set of genes it expresses and these expression programs are regulated by transcription factors.

Transcription factors are proteins that bind to deoxyribonucleic acid (DNA) in a sequence-specific manner to activate or repress gene expression. They can be classified into families based on the structure of their DNA-binding domain. There are estimated to be around 1400 transcription factors in humans, accounting for 6% of protein-coding genes (Vaquerizas *et al.*, 2009). This makes transcription factors one of the largest single classes of proteins to be encoded in the human genome (Young, 2011). The most common structural families in humans are the zinc finger, homeodomain and helix-loop-helix families, which together comprise more than 80% of all transcription factors (Vaquerizas *et al.*, 2009). However, many transcription factors have not been functionally characterised. Three single transcription factors (p53, oestrogen receptor and c-Fos), for example, account for more publications than all other transcription factors combined (Vaquerizas *et al.*, 2009). This demonstrates that there is still much to be learnt about the specific functions of individual transcription factors.

More complex organisms have increasingly complex regulation of gene expression. Humans, for example, have more transcription factors than yeast (both in absolute terms and as a ratio of protein-coding genes) (Levine and Tjian, 2003). In bacteria that do not need to respond to changing environments (such as the parasite *Rickettsia*), transcription factors account for <1% of the genome, in keeping with the role of transcription factors in coupling environmental stimuli and gene expression (Seshasayee *et al.*, 2011). Other examples of increasing complexity in higher organisms are the diversification of regulatory elements (for example, enhancers), the expansion of core transcriptional machinery components, and the binding of

transcription factors to diverse sets of co-factors (Levine and Tjian, 2003). These increasingly sophisticated mechanisms of regulation have been underpinned by the genetic diversification of transcription factors over time, associated with alterations in activity and specificity. Various evolutionary events have driven this diversification, including *de novo* development of functional domains, adaptation of ancestral transcription factors, gene duplications and lineage-specific losses (Weirauch and Hughes, 2011). The expanded gene regulatory capabilities of the diversified transcriptional machinery have in turn provided the foundation for more complex physiology, such as embryogenesis and multicellularity (Weirauch and Hughes, 2011). Transcription factors, in fact, still represent one of the most rapidly evolving classes of proteins today (Bustamante *et al.*, 2005).

As well as driving evolution and organism complexity, alterations in gene regulatory mechanisms can cause disease. Misregulation of gene expression is associated with a number of diseases, including developmental disorders, autoimmune diseases and cancer (Lee and Young, 2013). The mechanisms of altered regulation may include inappropriate expression, structure or function of a transcriptional regulator or mutation of a regulatory element (as selected examples). In the OMIM (Online Mendelian Inheritance in Man) database, there are 164 transcription factors that are directly responsible for 277 monogenic inherited disorders. Transcription factors are also commonly identified as oncogenes and tumour suppressor genes in cancer (Vaquerizas *et al.*, 2009). Understanding the mechanisms by which transcription factors regulate gene expression, and in turn how transcription factors are themselves regulated, is therefore a fundamental step in the development of treatments for these diseases.

### **Targeting Transcription Factors Therapeutically**

As the convergence point for multiple signalling pathway, transcription factors integrate a range of internal and external signals through mechanisms such as ligand binding, post-translational modifications, and interactions with other proteins and non-coding RNAs. In cancer, dysregulated transcription factors act as regulatory “hubs”, integrating oncogenic signals, and translating them into gene expression programs that underpin many of the hallmarks of cancer (Hanahan and Weinberg, 2011; Johnston and Carroll, 2015). Directly targeting abnormal gene expression by altering the activity of transcriptional regulators is therefore a compelling strategy in cancer treatment. Furthermore, the direct targeting of transcriptional regulators provides far fewer potential routes to therapeutic resistance compared to the indirect targeting of

upstream pathways. This is because signalling pathways are characterised by extensive redundancy and cross-talk, which facilitate the bypass of an inhibited component (Gonda and Ramsay, 2015).

Despite these advantages, however, transcription factors have been traditionally viewed as “undruggable”. One reason for this is that most transcription factors lack a clearly defined structural pocket for interactions with ligands or enzymatic substrates (a notable exception being the nuclear hormone receptors). Instead, the interaction interface between transcription factors and DNA is generally large and flat, making pharmacological targeting of this interaction with small molecules difficult (Lazo and Sharlow, 2016). Recent advances in pharmacology, computer modelling, and molecular biology, however, are challenging this paradigm, and transcriptional regulators are emerging as increasingly viable therapeutic targets.

There are a number of emerging strategies that can be used to therapeutically target transcription factors, many of which take advantage of the normal processes that regulate transcription factor activity. These strategies include: (1) Inhibition of protein-protein interactions (for example, with other transcription factors or co-factors), (2) Inhibition of enzymatic regulators (for example, kinases that directly phosphorylate the transcription factor), (3) Inhibition of DNA-protein interactions (for example, through the use of decoy oligonucleotides, which sequester transcription factors and prevent them from binding to DNA), (4) Targeting RNA degradation (for example, through the use of small interfering RNA (siRNA) or antisense oligonucleotides), and (5) Targeting proteins involved in epigenetic regulation, thereby affecting the accessibility of transcription factor binding (for example, histone modifying enzymes, chromatin remodellers, and DNA methyltransferases) (Johnston and Carroll, 2015; Yan and Higgins, 2013). Importantly, a number of these strategies do not rely on traditional small molecule approaches but instead utilise an evolving class of drugs known as biologics, including peptides, antibodies, and modified nucleic acids (Lazo and Sharlow, 2016).

The MYC proto-oncogene basic helix-loop-helix transcription factor (MYC) is an excellent example of a cancer-associated transcription factor that is being investigated as a therapeutic target. Several of the above strategies have been explored for MYC, including small molecule inhibitors of MYC dimerisation (essential for MYC transcriptional activity), a dominant negative peptide (Omomyc, which sequesters MYC and prevents dimerisation), and a lipid nanoparticle formulation of MYC-targeting siRNA (recently tested in a phase I clinical trial) (reviewed in Lazo and Sharlow, 2016).

Members of the E26 transforming sequence (ETS) transcription factor family are also important targets of drug development, due to their frequent dysregulation in many cancer types. A promising strategy is the inhibition of ETS factor protein-protein interactions, which are important in facilitating specific gene regulation by different ETS family members. One example is the EWS-FLI1 fusion protein small molecule inhibitor (YK-4-279), which inhibits the interaction of EWS-FLI1 with RNA helicase A (Erkizan *et al.*, 2009). An oral formulation of YK-4-279 has recently been described, although no clinical trials have yet commenced (Lamhamedi-Cherradi *et al.*, 2015). Subsequent studies suggest that this inhibitor may also be effective in prostate cancers with ETS gene fusions (Rahim *et al.*, 2011).

The ultimate aim of therapies targeting transcriptional regulators is to modify the gene expression programs that contribute to cancer development and progression. Effective therapeutic targeting of transcription factors will depend on a detailed understanding of how transcription factors function, the mechanisms that regulate them, and how these processes go wrong in cancer.

### **ELF5 and Breast Cancer**

One transcription factor with emerging roles in cancer is the E26 transforming sequence (ETS) factor E74-like factor 5 (ELF5). ELF5 is a master regulator of cell fate that guides the development of cell lineages in the placenta, lung, and mammary gland. In the mammary epithelium, ELF5 expression drives the development of the oestrogen receptor (ER)-negative alveolar cells that will become milk-producing cells in response to the hormonal cues of pregnancy. The balance between ER and ELF5 transcriptional activity in the mammary epithelial cell is therefore a fundamental determinant of cell fate (Gallego-Ortega *et al.*, 2013).

Similarly, in breast cancer, ELF5 expression is increased in the basal-like (mostly ER-negative) subtype, and is important for the maintenance of the gene expression programs that define this subtype. In addition, ELF5 is essential for the ongoing growth of cells in culture that have developed resistance to tamoxifen, one of the most commonly used endocrine therapies for ER-positive breast cancer (Kalyuga *et al.*, 2012). However, the molecular mechanisms by which ELF5 functions in these oestrogen-independent contexts are not well understood. Importantly, in the clinical setting, both ER-negative basal-like and ER-positive endocrine-resistant breast cancer are associated with poor prognosis. There is therefore a need for targeted treatments for these groups of patients, as well as biomarkers that can help predict the response



to treatments such as tamoxifen.

## **Thesis overview**

The aim of this thesis is to investigate how the lineage-defining transcription factor ELF5 functions in breast cancer, and what additional factors regulate these functions. This will facilitate the development of biomarkers and targeted treatments for basal-like and endocrine-resistant breast cancers. Specifically, four main areas will be explored:

1. The expression of *ELF5* at the transcript variant level in normal tissues and cancer;
2. The phenotypic and transcriptomic effects of modifying ELF5 expression in breast cancer cells;
3. The interactions with other proteins that contribute to these effects, and
4. The potential therapeutic applications of ELF5 transcriptional effects and protein interactions.

The background information in Chapter 1 is organised into five main parts. Firstly, general concepts related to how transcription factors function will be introduced (1.1), followed by a discussion of the additional factors that regulate these functions (1.2). The importance of these functional and regulatory mechanisms lies in their potential to be therapeutically harnessed to modulate transcription factor activity. Part 1.3 will provide background information on breast cancer, which is currently the most frequently diagnosed cancer in Australian women. The intrinsic subtypes of breast cancer, and the relationship between these subtypes and normal mammary development, will be discussed. Part 1.4 will focus on transcriptional dysregulation in cancer, using the ETS transcription factor family and ELF5 as examples. Finally, the emerging relationship between DNA repair and transcription will be explored in Part 1.5, arising from the discovery that ELF5 interacts with the DNA-dependent protein kinase catalytic sub-unit (DNA-PKcs, Chapter 5). This discussion will introduce DNA repair proteins as a novel class of transcriptional regulators, directly relevant to cancer development, progression, and treatment.

In Chapter 2, the materials and methods used in the experimental work are described.

Chapter 3 is a comprehensive analysis of ELF5 transcript variant expression in more than 6,000 normal tissue and cancer samples from The Cancer Genome Atlas (TCGA). Early studies of ELF5 described tissue-specific differences in transcript variant expression but recent studies have not distinguished between variants or have used a

single protein isoform for over-expression studies. This chapter investigates how ELF5 transcript variant expression is altered in cancer compared to normal tissues. It also defines the most highly expressed transcript variant in the normal breast and breast cancer, forming the basis for the ELF5 inducible cell line models used in subsequent chapters. Finally, the functional effects of increased expression of various ELF5 isoforms are explored in breast cancer cell lines, providing unique insights into the transcriptional functions of ELF5 and the role of the Pointed domain.

In Chapter 4, the effects of increased ELF5 expression in a luminal breast cancer context are investigated using next-generation sequencing technology. ELF5 reduces the oestrogen sensitivity of luminal breast cancer cells, reminiscent of its developmental role in promoting an ER-negative mammary epithelial cell fate. This may be important in the development of anti-oestrogen resistance in ER-positive breast cancer; however, the mechanisms by which this transcriptional rewiring occurs are not completely understood. Both RNA-sequencing and chromatin immunoprecipitation sequencing (ChIP-seq) are used to investigate the transcriptomic effects of increased ELF5 expression in ER-positive MCF7 breast cancer cells. The effects of increased ELF5 expression on the genomic binding of the ER pioneer factor FOXA1 are also investigated, representing a novel potential mechanism for modulation of the endocrine response.

Finally, Chapter 5 explores the protein interactions of ELF5 on chromatin using rapid immunoprecipitation of endogenous protein (RIME). Despite the importance of co-operative interactions to specific ETS factor function, no ELF5-interacting proteins have yet been identified in human breast or breast cancer cells. The investigations in this chapter validate the interaction of ELF5 with DNA-PKcs. A new transcriptional network between ELF5, ER and DNA-PKcs is proposed, with important potential implications for the use of DNA-PKcs inhibitors in breast cancer treatment.

Each results chapter includes a detailed discussion of the findings and their potential therapeutic implications. The final chapter (Chapter 6) will summarise the conclusions arising from these discussions, along with directions for future research.

## 1.1 How Transcription Factors Function

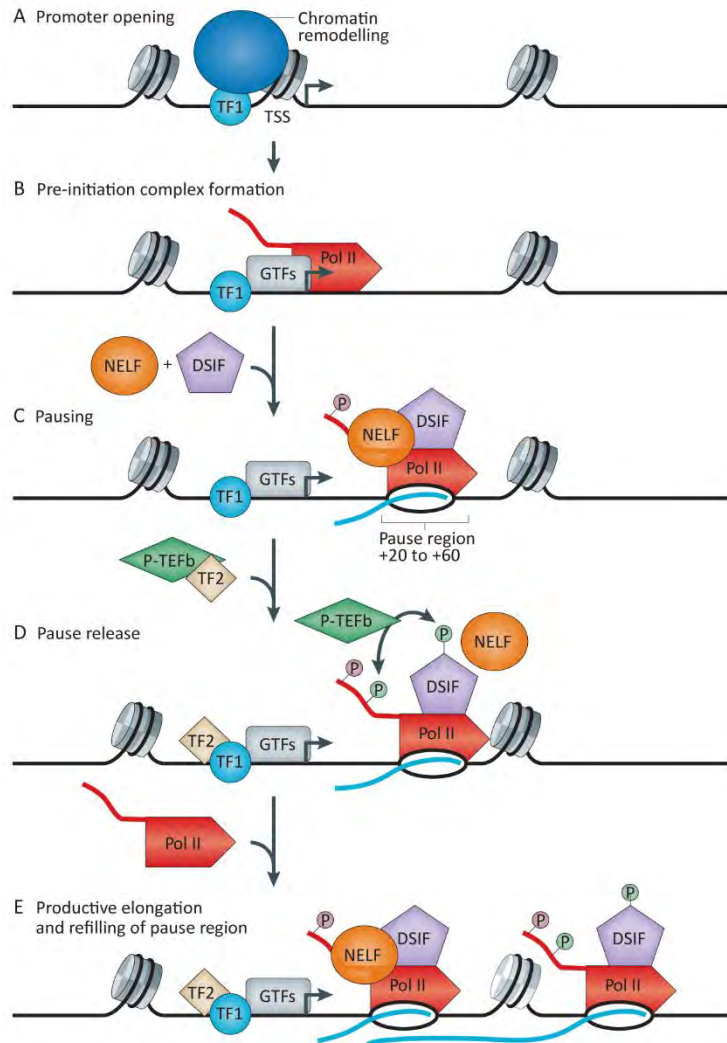
The basic principles of gene regulation were established by Jacob and Monod in 1961 (reviewed in Young, 2011), with the discovery that the *Escherichia coli* (*E. coli*) lac operon was controlled by specific repressors and activators that were bound to regulatory DNA sequences in the proximal gene promoter. In higher organisms, the regulation of gene expression has added levels of complexity. DNA, for example, is packaged into nucleosomes, restricting transcription factor access. In addition, transcription factors often bind to regulatory sites that are many thousands of base pairs away from the transcription start site. The following section will discuss the mechanisms by which transcription factors are able to regulate gene expression in eukaryotes.

### General Process of Transcription

Transcription is the process of synthesising a ribonucleic acid (RNA) molecule from a complementary strand of DNA. For protein-coding genes, the resulting messenger RNA (mRNA) molecule may then be translated into amino acids by the cellular machinery to form a protein. Transcription in humans is catalysed by three RNA polymerase enzymes. RNA polymerase II produces mRNA and some small nuclear RNAs, while RNA polymerases I and III produce primarily ribosomal RNA (rRNA) and transfer RNA (tRNA) (Cramer *et al.*, 2008).

The mRNA-producing RNA polymerase II (Pol II) is a large multi-sub-unit complex. It is composed of 12 sub-units (RPB1-12), with the largest two subunits forming opposite sides of an enzymatic cleft. The sub-units on each side of the cleft form “jaws” which grip the downstream DNA. The active site within the cleft contains magnesium ions and catalyses the addition of nucleotides to the growing RNA chain (Cramer *et al.*, 2008; Cramer *et al.*, 2000). The C-terminal domain of the RPB1 sub-unit (Pol II CTD) is flexibly linked to the core enzyme and contains 52 repeats of a 7-peptide sequence (Barrero and Malik, 2013). Phosphorylation of the CTD occurs during transcription and provides a platform for recruitment of additional transcriptional and RNA-processing factors (Zhou *et al.*, 2012).

The process of transcription consists of a number of steps (Figure 1.1). Transcription is initiated by the binding of a sequence-specific transcription factor to a regulatory element (promoter or enhancer). Chromatin remodellers may be recruited by the transcription factor, resulting in increased accessibility of promoter DNA (Figure 1.1A).



**Figure 1.1: The general process of transcription consists of several stages**

**Figure reproduced (and caption modified) with permission from Nature Publishing Group (Adelman and Lis, 2012).**

The promoter region is shown with the transcription start site (TSS) labelled with an arrow. Nucleosomes are depicted in grey and RNA polymerase II (Pol II) is illustrated as a red rocket. The nascent RNA transcript is shown in blue. (A) Binding of a sequence-specific transcription factor (TF1) recruits chromatin remodellers (dark blue), resulting in increased accessibility of promoter DNA. (B) Pol II and general transcription factors (GTFs) bind to the core promoter to form the pre-initiation complex (PIC), aided by the binding of sequence-specific transcription factors. (C) Transcription commences and Pol II clears the promoter region, pausing after 20-60 bp (promoter-proximal pausing). Pausing is controlled by the pause factors DSIF and NELF. The paused Pol II is phosphorylated on the C-terminal domain (serine 5, shown in pink). (D) Pause release requires the action of P-TEFb, which may be recruited by sequence-specific transcription factors (TF2). P-TEFb phosphorylates Pol II on the C-terminal domain (serine 2, shown in green) as well as DSIF and NELF. Phosphorylated NELF dissociates while phosphorylated DSIF is converted to a positive

elongation factor and remains associated with Pol II. (E) Pol II proceeds to elongation, moving down the DNA template. A second Pol II complex enters the pause site, allowing for efficient RNA production. DSIF, DRB sensitivity inducing factor; NELF, negative elongation factor; P-TEFb, positive elongation transcription factor b.

RNA Pol II and general transcription factors then bind to the core promoter region, forming the pre-initiation complex (PIC) (Figure 1.1B) (Adelman and Lis, 2012). The core promoter is a region spanning about 40 base pairs (bp) upstream and downstream from the transcription start site (TSS) containing common DNA elements (such as the TATA box) to which general transcription factors can bind (Goodrich and Tjian, 2010). General transcription factors, which include transcription factors (TF) IIA, B, D, E, F and H, help to position and orient RNA Pol II on the core promoter (Barrero and Malik, 2013). The co-activator Mediator is also an essential component of the PIC (Allen and Taatjes, 2015).

TFIID is usually the first general transcription factor to bind to the core promoter. The TFIID complex consists of the TATA-binding protein (TBP) and 13-14 TBP-associated factors (TAFs). The TBP sub-unit interacts with the TATA box (consensus sequence TATAA), while other TAFs can interact with sequences such as the initiator element (Inr) or the downstream promoter element (DPE) (Goodrich and Tjian, 2010). TFIIA and TFIIB help to stabilise the binding of TFIID, with TFIIB forming the main bridge between the promoter DNA and RNA Pol II (Hantsche and Cramer, 2016). TFIIIE and TFIIH are the last of the general transcription factors to bind. TFIIH has helicase / DNA translocase activity and contributes to promoter “melting” (process by which DNA around the TSS is partially unwound) (Barrero and Malik, 2013; Hantsche and Cramer, 2016). TFIIH also phosphorylates the CTD at serine 5, an essential step in allowing Pol II to clear the promoter and initiate transcription (Allen and Taatjes, 2015).

Once transcription is initiated, it proceeds for a short distance (20-60 bp) and then pauses (promoter-proximal pausing, Figure 1.1C). Proposed functions of pausing include maintenance of chromatin in an open state, facilitation of rapid gene induction in response to signals and recruitment of RNA-processing factors (Adelman and Lis, 2012). Pausing is controlled by the pause factors DRB sensitivity inducing factor (DSIF) and negative elongation factor (NELF). Pause release (Figure 1.1D) requires the action of positive elongation transcription factor b (P-TEFb), which is a complex containing cyclin-dependent kinase 9 (CDK9) and cyclin T1 or T2 (Zhou *et al.*, 2012). P-TEFb phosphorylates the RNA Pol II CTD (serine 2) as well as the pause factors DSIF and NELF. P-TEFb may be targeted to promoters for pause release by sequence-specific

transcription factors or by the bromodomain-containing protein 4 (BRD4), which interacts with acetylated histones (Adelman and Lis, 2012; Heinz *et al.*, 2015). Phosphorylation of NELF causes it to dissociate, while phosphorylation of DSIF converts it into an elongation-promoting factor. Phosphorylated DSIF remains bound to RNA Pol II as elongation proceeds (Adelman and Lis, 2012).

Additional proteins are recruited to form the elongation complex and RNA Pol II moves along the DNA strand (Figure 1.1E). Nucleotides are added to the RNA chain by the active enzymatic site of Pol II at a rate of about 3.8 kilobases (kb) per minute (Zhou *et al.*, 2012). When Pol II reaches the polyadenylation (poly(A)) signal, the RNA chain is cleaved and a poly(A) tail added. Dissociation of Pol II from the DNA occurs further downstream and may involve structural rearrangements due to the binding of RNA-processing factors and/or nuclease action (Hantsche and Cramer, 2016).

In the cellular context, transcription is influenced by many variables, including the binding of sequence-specific transcription factors, local chromatin structure and the expression of tissue-specific general transcription factors and co-factors. Sequence-specific transcription factors, binding to promoter or enhancer elements, are able to regulate the transcriptional process at a number of stages, including PIC recruitment, initiation and pause release (Fuda *et al.*, 2009). This will be discussed in more detail below. In addition, it is becoming increasingly recognised that the core transcriptional machinery itself is involved in the regulation of transcription. Different tissues, for example, can express different TFIID sub-units, facilitating the dynamic assembly of non-canonical pre-initiation complexes, which can have tissue-specific functions (Goodrich and Tjian, 2010).

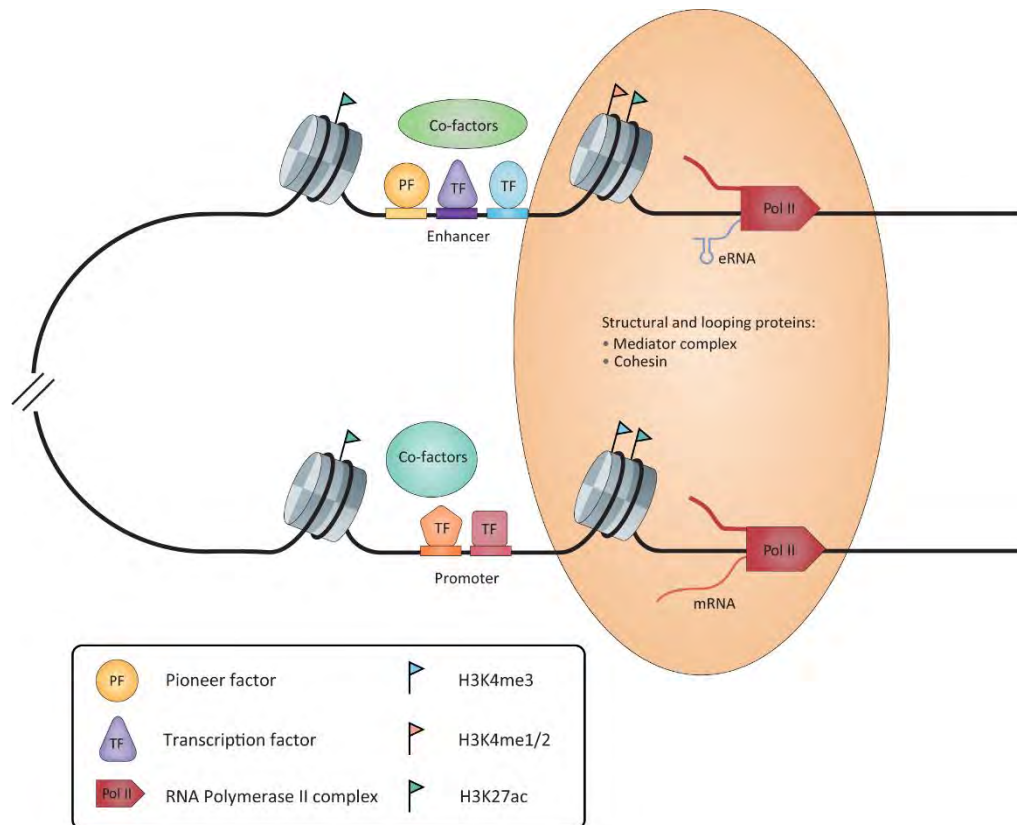
### **Interactions with the Core Transcriptional Machinery**

At the most basic level, transcription factors act by recruiting and interacting with the core transcriptional machinery (either directly or indirectly) to activate or repress gene transcription. Most transcription factors are thought to regulate transcriptional initiation. The transcription factors p53, SP1 and JUN, for example, can bind to various TFIID sub-units, indicating they may directly recruit the pre-initiation complex (PIC) to transcriptional sites (Liu *et al.*, 2009). This is further supported by DNase “footprinting” experiments, in which SP1 motifs were found to be enriched within the 50 bp PIC binding region (“footprint”) of gene promoters (Neph *et al.*, 2012).

Transcription factors also interact indirectly with the core transcriptional machinery through co-factors. Mediator, for example, is a large multi-sub-unit complex that

interacts with Pol II as well as various sequence-specific transcription factors. The general role of Mediator is to integrate regulatory signals from DNA-bound transcription factors to the transcriptional machinery, although the exact mechanisms by which Mediator regulates Pol II activity are not well understood (Allen and Taatjes, 2015). Mediator also contributes to the formation of DNA loops that bring enhancer and promoter regions into close proximity (Figure 1.2). Co-factors recruited by transcription factors may also alter chromatin structure, increasing accessibility of DNA to the transcriptional machinery, or modify histones (Farnham, 2009). Specific histone modifications have been shown to directly recruit the transcriptional machinery. The plant homeodomain (PHD) finger of the TFIID sub-unit TAF3, for example, can directly bind to histone H3 with lysine 4 trimethylation (H3K4me3), a modification associated with promoter regions and transcription start sites (Vermeulen *et al.*, 2007).

Transcription factors can also regulate other stages of transcription, including promoter-proximal pausing and elongation. The transcription factor MYC, for example, regulates pause release by binding to P-TEFb, which is then able to phosphorylate DSIF and NELF to allow transition into elongation (Rahl *et al.*, 2010). MYC has also been found to cause increased transcription of most actively expressed genes, acting as a transcriptional amplifier through its positive effect on elongation (Lin *et al.*, 2012; Nie *et al.*, 2012). The autoimmune regulator (AIRE) also regulates pause release in the thymus to allow expression of a wide range of ectopic genes important in the development of immunological self-tolerance (Giraud *et al.*, 2012). Thus, through both direct and indirect interactions with the core transcriptional machinery, transcription factors can regulate multiple stages of transcription, including initiation, promoter-proximal pausing and elongation.



**Figure 1.2: Multiple mechanisms of action of eukaryotic transcription factors**

*Based on figure from (Heinz et al., 2015)*

Transcription factors bind to specific DNA sequences within the gene promoter or distal enhancers. Binding of transcription factors results in the recruitment of co-factors including histone modifiers and chromatin remodellers. Together these factors increase the accessibility of DNA that is packaged within nucleosomes (pictured as grey cylinders). Transcription factors can interact with the core transcriptional machinery (Pol II and associated factors) to regulate transcriptional initiation, promoter-proximal pausing and elongation. Many transcription factors can perform this function from sites that are located many thousands of base pairs away (enhancers). Enhancers typically require the co-operative action of multiple transcription factors to regulate gene expression, including pioneer factors (which contribute to cell-specific enhancer selection) and additional transcription factors responsive to regulatory signals. Enhancers can interact with gene promoters to regulate transcription by DNA looping, facilitated by proteins such as Mediator and Cohesin. Active enhancers are characterised by histone modifications such as histone H3 lysine 4 monomethylation (H3K4me1), dimethylation (H3K4me2) and H3 lysine 27 acetylation (H3K27ac), while gene promoters have a relative enrichment for histone H3 lysine 4 trimethylation (H3K4me3). eRNA, enhancer RNA; mRNA, messenger RNA.



## Enhancers

Many transcription factors act by binding to distal regulatory elements called enhancers. In general, enhancers are a few hundred base pairs in length, are located at a distance from the regulated gene (hundreds up to millions of bases away) and can regulate gene expression when bound by sequence-specific transcription factors. Enhancers typically contain binding sites for multiple transcription factors, which can act co-operatively to regulate gene expression. Enhancers are vital in the regulation of cell-type-specific gene expression (Pott and Lieb, 2015) and many cell-type-specific genes are regulated by multiple enhancers (Young, 2011).

The first enhancer was identified in the SV40 virus when it was found that inclusion of SV40 DNA in a  $\beta$ -globin expression vector markedly increased transcription from the vector in cell lines (Banerji *et al.*, 1981). The enhancing activity was specific to a region of the SV40 genome containing two 72 bp repeats and these sequences could function in multiple positions and orientations within the vector. Later studies identified binding sites for several sequence-specific transcription factors in the SV40 enhancer (reviewed in Levine, 2010). Thus, several features of enhancers were defined by these studies, including the ability to augment transcription, a lack of strict positional dependence relative to the promoter sequence and the regulation by multiple transcription factors.

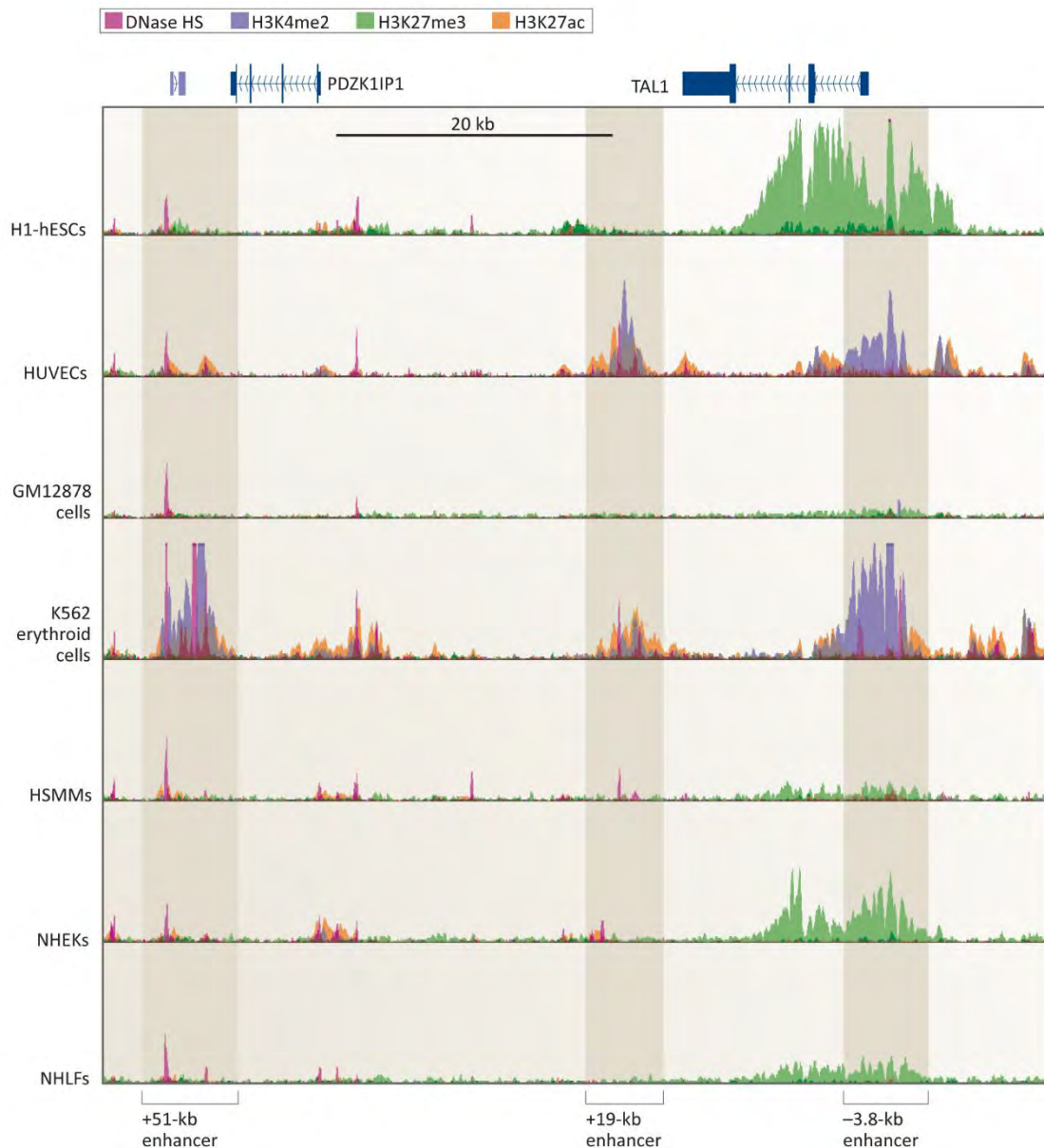
More recent genome-wide studies have identified additional features of enhancers. These include an enrichment for histone H3 lysine 4 monomethylation and dimethylation (H3K4me1 or 2) and a depletion of histone H3 lysine 4 trimethylation (H3K4me3) compared with promoters. Enhancers are also contained in regions of accessible chromatin, as demonstrated by sensitivity to digestion by DNase I (DNase I hypersensitivity). An active enhancer is characterised by the additional presence of histone H3 lysine 27 acetylation (H3K27ac), as well as RNA Pol II and transcriptional co-factors such as p300 (Heinz *et al.*, 2015). These features are demonstrated in Figure 1.3 for the T-cell acute lymphocytic leukaemia 1 (TAL1) gene in several cell lines (adapted from Heinz *et al.*, 2015).

Based on these features, the Encyclopedia of DNA Elements (ENCODE) project has identified approximately 400,000 putative enhancers in the human genome (Dunham *et al.*, 2012), although the vast majority have not been functionally characterised (Pott and Lieb, 2015). This means that the number of regulatory elements in the genome far exceeds the number of protein-coding genes. In a given cell type, 10,000-150,000 enhancers are estimated to be active (Pott and Lieb, 2015). Most DNase I

hypersensitive sites (DHSs, which indicate accessible chromatin) were located in intronic and intergenic regions, with only 5% of DHSs located within 2.5 kb of a transcription start site at gene promoters. DHSs within promoters tended to be found in multiple cell types, while distal DHSs (largely enhancers) were usually cell-specific (Thurman *et al.*, 2012). This provides additional evidence for the importance of distal elements in the regulation of cell-type-specific gene expression.

The activation of an enhancer typically involves the binding of multiple transcription factors, including pioneer factors and signal-responsive transcription factors (Figure 1.2). In this way, enhancers can integrate multiple signals, including intrinsic cellular factors (for example, the set of transcription factors expressed) and external environmental cues (Calo and Wysocka, 2013; Heinz *et al.*, 2015). Different transcription factors may act co-operatively to access enhancer chromatin, effectively “out-competing” the nucleosome for binding to DNA. Enhancers can also be made accessible by the action of pioneer factors, which have the unique ability to bind compacted chromatin (reviewed in Zaret and Carroll, 2011). Bound transcription factors recruit co-factors such as histone modifiers and chromatin remodellers to the active enhancer, which then interacts with gene promoters through DNA looping (Calo and Wysocka, 2013), as shown in Figure 1.2. DNA looping is facilitated by proteins such as Mediator and Cohesin (Kagey *et al.*, 2010).

New techniques such as chromosome conformation capture (3C and its derivatives) and Pol II chromatin interaction analysis with paired-end tag sequencing (ChIA-PET) have allowed interactions between enhancers and promoters to be mapped. These methods have also revealed a perhaps unanticipated complexity in regulatory element interactions. In the ENCODE pilot project regions (about 1% of the human genome), 50% of gene promoters engaged in at least one long-range interaction, with some interacting with as many as 20 distal elements. Similarly, many distal fragments interacted with more than promoter (Sanyal *et al.*, 2012). This is likely to reflect both complex combinatorial gene regulation as well as cell-to-cell variation. Only 7% of the long-range interactions identified by ENCODE were between a distal fragment and the nearest transcriptional start site and some interactions were found to skip over intervening genes and/or CCCTC-binding factor (CTCF)-bound sites (Sanyal *et al.*, 2012). This demonstrates that the mechanisms by which enhancers are directed to their target genes are not currently well-understood.



**Figure 1.3: Cell-type-specific enhancers have characteristic chromatin features**

**Figure reproduced (and caption modified) with permission from Nature Publishing Group (Heinz et al., 2015)**

Genomic features of a 60 kb region of human chromosome 1 in multiple cell lines, centred around the T-cell acute lymphocytic leukaemia 1 (TAL1) gene. Encyclopedia of DNA Elements (ENCODE) data are shown (overlaid) for DNase I hypersensitivity (marking regions of accessible chromatin, pink) and chromatin immunoprecipitation followed by sequencing (ChIP-seq) for several histone modifications. Histone H3 lysine 4 dimethylation (purple) and histone H3 lysine 27 acetylation (orange) ChIP-seq signals indicate regions associated with active enhancers and histone H3 lysine 27 trimethylation (green) ChIP-seq signal indicates regions associated with gene repression. TAL1 is regulated by 3 distinct enhancers (shaded), located at -3.8-kb, +19-kb and +51-kb from the transcription start site. In endothelial cells (HUVECs), 2 of the enhancers (-3.8-kb and +19-kb) are active, as shown

by increased DNase hypersensitivity and the presence of H3K4me2 and H3K27ac. In K562 erythroid cells, the +51-kb enhancer is also active, demonstrating the cell-type-specificity of enhancers. In cell types where TAL1 is not expressed, there is an absence of H3K4me2 and H3K27ac and variable levels of the repressive modification H3K27me3. H1-hESCs, human embryonic stem cells; HUVECs, human umbilical vein endothelial cells; GM12878 cells, human lymphoblasts; HSMs, human skeletal muscle myoblasts; NHEKs, normal human epidermal keratinocyte; NHLFs, normal human lung fibroblast; PDZK1IP1, PDZK1-interacting protein 1.

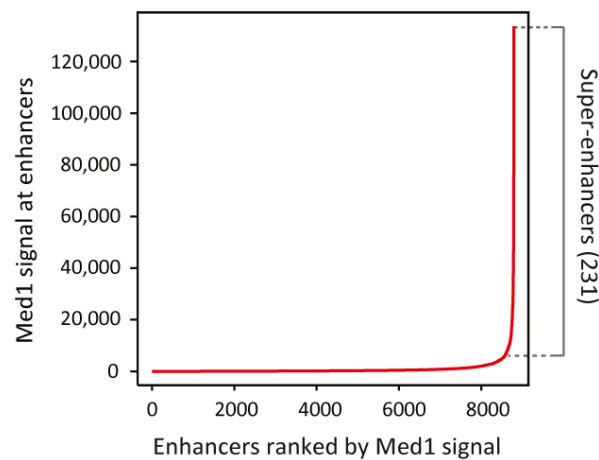
## **Super-Enhancers**

A related concept is “super-enhancers”, described independently by two groups in 2013 (Loven *et al.*, 2013; Whyte *et al.*, 2013). Super-enhancers are essentially clusters of enhancers spanning large genomic regions. They are characterised by high levels of sequence-specific transcription factors (typically “master regulators” of the cell lineage), as well as Mediator, H3K4me1 and H3K27ac. Super-enhancers are able to stimulate higher transcriptional activity than typical enhancers and largely regulate genes that control cell identity. An ongoing debate is whether the constituent enhancers act independently (as strong and robust regulators) or whether their clustering confers an additional benefit that exceeds the sum of their parts (Dukler *et al.*, 2016; Pott and Lieb, 2015).

In murine embryonic stem cells (ESCs), 231 super-enhancers were identified using chromatin immunoprecipitation followed by sequencing (ChIP-seq) for the key ESC transcription factors Oct4, Sox2 and Nanog, as well as the Mediator sub-unit Med1 (Whyte *et al.*, 2013). The method used to define super-enhancers in this study is illustrated in Figure 1.4. The median size of the 231 super-enhancers was approximately 9000 bp, compared to just 700 bp for typical enhancers. Motifs for Oct4, Sox2 and Nanog were identified in the super-enhancers, as expected, along with motifs for Klf4 and Esrrb. Interestingly, each of the master transcription factors (Oct4, Sox2, Nanog, Klf4 and Esrrb) were themselves driven by a super-enhancer, setting up auto-regulatory loops. This is likely to be an important mechanism by which lineage-specific transcription factors establish robust gene expression programs. Analysis of the transcription factors regulated by super-enhancers could also provide a method for identifying novel master regulators in different cell types (Hnisz *et al.*, 2013).

Since their initial characterisation, super-enhancers have been described in many cell types, for example adipocytes (Siersbaek *et al.*, 2014), hair follicle stem cells (Adam *et al.*, 2015) and mammary epithelial cells (Shin *et al.*, 2016). Various methods have been

used to identify super-enhancers (reviewed in Niederriter *et al.*, 2015). Hnisz *et al.* (2013), for example, used H3K27ac ChIP-seq data alone to identify and rank enhancers in a panel of 86 human cells and tissues. One advantage of using H3K27ac or Mediator ChIP-seq data alone is that it allows identification of super-enhancers even when the master regulatory transcription factors are unknown or where no transcription factor ChIP-seq is available. The use of different methods does, however, make comparison of studies more difficult.



**Figure 1.4: Super-enhancers can be defined based on Mediator binding signal**

**Figure reproduced (and caption modified) with permission from Elsevier (Whyte *et al.*, 2013)**

ChIP-seq enhancer peaks were identified for the embryonic stem cell transcription factors Oct4, Sox2 and Nanog. Peaks within 12.5 kb were combined (or “stitched”) and all enhancers were then ranked by Med1 ChIP-seq signal. This produced a curve with a clear inflection point, where the Med1 signal began to rapidly increase. All enhancers above this geometrically-defined point were classified as super-enhancers.

As indicated above, the biological relevance of the term “super-enhancer” is controversial. Pott and Lieb (2015) argue that there is no evidence to support the concept that super-enhancers are a novel regulatory mechanism but instead represent highly-bound, highly-active tissue-specific enhancers. The definition of a super-enhancer, for example, is solely based on ChIP-seq features and there is no functional element to the selection of the cut-off used to separate typical enhancers from super-enhancers. Functional characterisation, they argue, is essential to determining whether a super-enhancer functions as a single unit that is indeed greater than the sum of its parts.

Two recent studies have attempted to address this question by genetically altering constituent enhancers within super-enhancers. These studies each focused on a single murine super-enhancer, driving either the mammary Whey acidic protein (*Wap*) gene (Shin *et al.*, 2016) or the erythroid  $\alpha$ -globin gene (Hay *et al.*, 2016). Interestingly, the two studies reached different conclusions regarding the question of synergistic function. For the *Wap* super-enhancer, the three constituent enhancers were found to function hierarchically; mutation of multiple transcription factor binding sites in the proximal enhancer, for example, prevented the establishment of the additional two enhancers in the normal temporal manner during pregnancy, thereby disabling the entire super-enhancer. In the case of the  $\alpha$ -globin super-enhancer, the authors concluded that each constituent enhancer appeared to act independently and additively, indicating an absence of synergistic function. Dukler *et al* (2016) argue that a simple generalised linear mathematical model, that does not require interactions between individual enhancers, explains the data generated by Shin and Hay. With the data currently available, they argue, super-enhancers cannot confidently be declared to function as units that exceed the sum of their parts.

Regardless of whether super-enhancers function as additive or synergistic functional units, what is clear from the above studies is that a subset of highly-active enhancers is essential for driving cell-type-specific gene expression programs. These clusters of enhancers act as binding platforms for lineage-specific master transcription factors, which themselves are regulated by super-enhancers in robust auto-regulatory loops.

### **DNA-Binding Specificity**

One of the first steps in the action of any transcription factor is the binding to DNA. Transcription factors interact with specific DNA sequences within the genome, which are usually 4-20 bp in length. Recognition of these sites occurs through physical interactions between amino acid side chains of the transcription factor and DNA nucleotides (“base readout”) as well as through sequence-dependent DNA structure (“shape readout”) (Slattery *et al.*, 2014). Hydrogen bonds form between amino acids of the DNA-binding domain in the transcription factor and the bases of the DNA molecule. Most transcription factor interactions occur within the major groove of DNA (Odom, 2011).

An optimal “cognate” binding sequence has been identified for many transcription factors, characterised by the lowest (most favourable) free energy of binding (Crocker *et al.*, 2016). Transcription factors bind to these cognate sites with high affinity. The

E26 transforming sequence (ETS) family member E74-like factor 5 (Elf5), for example, has a 13 bp *in vitro* consensus motif characterised by a core GGAA/T sequence that can be bound by all ETS family members (Choi and Sinha, 2006). The structure of the DNA-binding domain is the main determinant of sequence specificity. The alteration of a single amino acid residue in the DNA-binding domain of the ETS factor Elf1, for example, impairs the ability of this unique ETS factor to selectively bind GGAA core sequences (rather than GGAA/T) (Bosselut *et al.*, 1993).

However, transcription factors *in vivo* can bind to both high-affinity and low-affinity sequences. In fact, it has been shown that the highest-affinity sites are rarely occupied by eukaryotic transcription factors *in vivo* and that lower-affinity binding sites are frequently important for precise regulation of gene expression (reviewed in Crocker *et al.*, 2016). This means that eukaryotic transcription factors can recognise and bind to a range of sequences, which tend to be short and can therefore occur randomly throughout the genome. This raises the question of how transcription factors can act as specific regulators of gene expression. An equally puzzling question is how transcription factors that bind to highly similar sequences, for example the ETS family members, differentially regulate gene expression.

In addition, transcription factors can interact non-specifically with DNA. For the ETS family member ETS1, for example, functional DNA binding is characterised by the formation of hydrogen bonds and conformational changes, whereas non-specific interactions are mediated by electrostatic interactions between positively-charged amino acid chains and negatively charged DNA. These non-specific interactions have been shown to have a role, however, in allowing the transcription factor to slide along the DNA in order to rapidly scan the sequence for specific binding sites (Desjardins *et al.*, 2016).

DNA sequence specificity is an important mechanism of transcription factor action; however, it is not sufficient by itself to determine functional transcription factor binding in eukaryotes (Todeschini *et al.*, 2014; Wunderlich and Mirny, 2009). Transcription factors can potentially bind many genomic sites, although *in vivo* only a fraction of these sites are occupied. Additional mechanisms have evolved to increase the ability of transcription factors to specifically regulate gene expression. These include clustering of transcription factor binding sites (for example, in enhancers), co-operative transcription factor binding, interactions of transcription factors with co-factors, regulation of chromatin accessibility and enzymatic modifications of histones and DNA, all of which will be discussed in further detail below (Todeschini *et al.*, 2014).

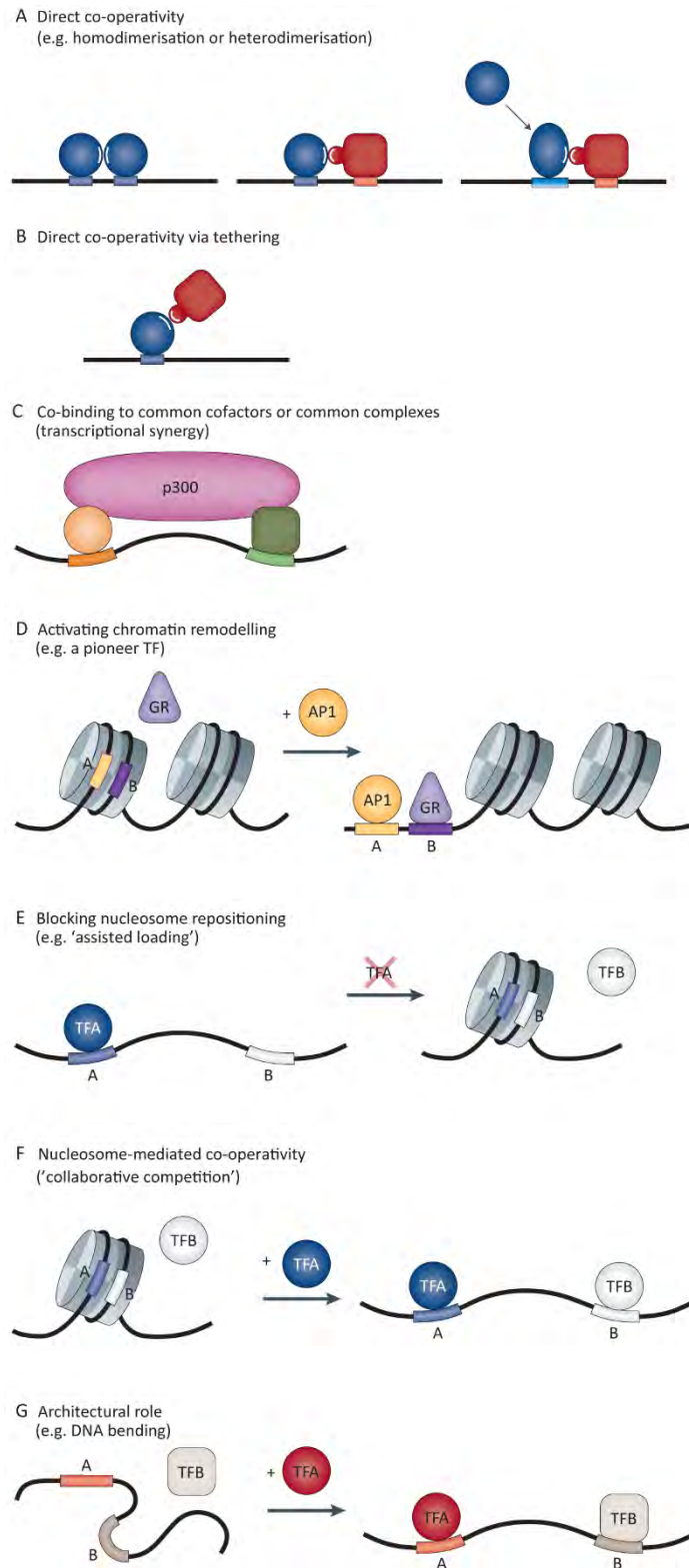
## Transcription Factor Co-operativity

An important principle of eukaryotic gene regulation is the co-operative action of multiple transcription factors (Figure 1.2). One of the earliest genome-wide binding studies in yeast demonstrated that multiple transcription factors frequently bind to the same promoter (Lee *et al.*, 2002). The binding of multiple transcription factors to the same regulatory region has also been seen in numerous ChIP-seq studies (for example, Gerstein *et al.*, 2012). The requirement for the binding of more than one transcription factor enables the integration of multiple signalling pathways, as well as an increase in regulatory specificity.

The mechanisms of transcription factor co-operation (shown in Figure 1.5) may be direct or indirect (Spitz and Furlong, 2012). Direct co-operativity involves protein-protein interactions between transcription factors and increases specificity by limiting binding to dimeric sites (Lelli *et al.*, 2012). One example is the obligate dimerisation of nuclear receptors such as oestrogen receptor (ER) (Figure 1.5A, left). Ligand-bound ER molecules bind to DNA in a head-to-head orientation and interactions between the ligand-binding domains (LBDs) assist in the correct positioning of the molecules on the DNA (Pardee *et al.*, 2011). The ETS transcription factor ETS proto-oncogene 1 (ETS1) can also bind to palindromic DNA sequences as a homodimer. In this case, the mechanism of co-operativity involves alteration of the structure of the second ETS1 molecule, resulting in the relief of DNA-binding auto-inhibition (Baillat *et al.*, 2002).

Non-identical transcription factors may also directly co-operate, forming heterodimers on DNA (Figure 1.5A, middle). These transcription factors may be from the same or different structural families and a single transcription factor can often have multiple partners (Lelli *et al.*, 2012). ETS1 and Runt-related transcription factor 1 (RUNX1), for example, can form a heterodimer in haematopoietic cells. The two factors bind to a composite DNA sequence, resulting in the displacement of one of the auto-inhibitory modules of ETS1 (Hollenhorst *et al.*, 2009; Shrivastava *et al.*, 2014). The binding of ETS1 as a heterodimer is important in directing enhancer-specific ETS1 binding and function; in contrast, many of the proximal-promoter sites bound by ETS1 are also redundantly bound by other ETS factors (Hollenhorst *et al.*, 2009). Just like ETS1, the transcription factor RUNX1 can co-operate with other factors to mediate specific effects. Throughout the different stages of megakaryocyte development, for example, RUNX1 can co-operate with GATA binding protein 1 (GATA1), activator protein 1 (AP1) and ETS factors (Pencovich *et al.*, 2011).





**Figure 1.5: Mechanisms of transcription factor co-operativity may be direct or indirect**  
*Figure and caption adapted with permission from Nature Publishing Group (Spitz and Furlong, 2012)*

A and B illustrate mechanisms of direct transcription factor co-operativity, while C-F illustrate indirect mechanisms. (A) Transcription factors can directly co-operate by binding as

homodimers (left) or heterodimers (middle). A related concept is latent specificity; when co-bound with the red transcription factor, the blue transcription factor can bind to a low-affinity site (right). (B) Tethering of a second transcription factor (red) by binding of a transcription factor (blue) to DNA. (C) Indirect co-operativity may occur through the interaction with common co-factors, for example the histone acetyltransferase p300, the Mediator complex or repressors such as Groucho. (D) The binding of one transcription factor (activator protein 1, AP1) may initiate chromatin remodelling, thereby allowing a second transcription factor (glucocorticoid receptor, GR) to bind. (E) Rapid, dynamic exchange in the binding of two transcription factors (TFA, TFB) that recognise very similar sites can result in an increase in the binding of each transcription factor by enhancing chromatin accessibility (assisted loading). (F) In nucleosome-mediated co-operativity (or collaborative competition), multiple transcription factors (TFA, TFB) “out-compete” histones for binding to DNA, resulting in a local increase in chromatin accessibility. (G) A transcription factor (TFA) may induce local changes in DNA conformation, thereby enhancing the binding of additional factors (TFB).

A variation on direct co-operativity is a phenomenon called “latent specificity” (Figure 1.5A, right). This involves an increase in the affinity of a transcription factor to a non-canonical site when bound co-operatively with another factor. The inherent DNA-binding specificity of the individual factors is unaltered, as the transcription factor is unable to bind the sequence in the absence of its partner (Lelli *et al.*, 2012). ETS1, for example, can co-operatively bind with the paired box 5 (PAX5) transcription factor on the CD79a gene promoter in B cells. The CD79a promoter contains a non-canonical ETS binding site with a core GGAGG motif (non-canonical nucleotide underlined). DNA-bound PAX5 can shift the conformation of an ETS1 side-chain tyrosine residue, altering the contacts formed between ETS1 and DNA. This allows co-operative ETS1 binding to this normally low-affinity site (Garvie *et al.*, 2001). A recent *in vitro* study also identified 315 pairs of transcription factors that displayed co-operative binding, with 207 pairs (66%) demonstrating markedly different specificity (often in the form of composite sites) when binding together compared with binding individually (Jolma *et al.*, 2015). This indicates that latent specificity may be a widespread regulatory mechanism.

Another mechanism of direct co-operativity occurs via tethering (Figure 1.5B). In this case, a DNA-bound transcription factor interacts directly with another transcription factor that is not bound to DNA. Multiple transcription factors, for example, can tether oestrogen receptor alpha (ER) to DNA, including cAMP response element binding protein 1 (CREB1), progesterone receptor (PR) and AP1. ER tethered to DNA via these factors is able to regulate oestrogen-responsive gene transcription (Heldring *et al.*, 2011; Mohammed *et al.*, 2015).

Transcription factors can also co-operate through indirect mechanisms. Co-bound transcription factors, for example, may interact with common co-factors, including the histone acetyltransferase p300, the Mediator complex or repressors such as Groucho (Figure 1.5C). This may have a number of effects, for example, stabilising co-factor binding, increasing the affinity of transcription factors for their binding sites or increasing the residence time for each transcription factor on DNA (Spitz and Furlong, 2012).

A second indirect mechanism is the initiation of chromatin remodelling by one transcription factor, which then enables the binding of additional regulators (Figure 1.5D). An example of this is the co-operation between AP1 and glucocorticoid receptor (GR). Most GR binding occurs at sites that already contain accessible chromatin prior to hormone stimulation. This baseline accessibility is mediated by the binding of AP1 to these sites, thereby priming them for rapid activation. In the absence of AP1, GR is unable to access these sites to regulate gene transcription following glucocorticoid stimulation. The factors that direct GR to accessible chromatin may be cell-type-specific, as demonstrated by the enrichment of different motifs at GR binding sites in murine hepatocytes (AP1, SP1 and forkhead) compared to mammary epithelial cells (AP1 only) (Biddie *et al.*, 2011). Another example is the relationship between pioneer factor forkhead box A1 (FOXA1) and ER in the mammary gland, where almost all ER binding is dependent on the chromatin remodelling initiated by FOXA1 (Hurtado *et al.*, 2011). Importantly, neither of these co-operative events require the direct interaction of the two transcription factors.

Thirdly, transcription factors may indirectly co-operate through a mechanism known as assisted loading (Figure 1.5E). Two transcription factors that bind to very similar sites may intuitively be thought to inhibit each other's activity. In contrast, it has been shown that rapid, dynamic exchange in the binding of these two factors on DNA (in the range of a few seconds) can enhance the binding of both factors by transiently increasing chromatin accessibility (described as a "hit and run" mechanism). This has been shown for the binding of GR and an ER variant designed to bind to glucocorticoid response elements (GREs). The enhancement in ER binding driven by GR binding occurs only at *de novo* GR binding sites (that is, at those sites that do not have accessible chromatin prior to hormone stimulation) and is a result of GR-mediated recruitment of chromatin remodellers such as SWI/SNF (Voss *et al.*, 2011). This interaction between GR and ER also occurs at naturally-occurring GR and ER response elements in mouse mammary epithelial cells, with the binding of either factor able to enhance binding of the other as

a result of assisted loading (Miranda *et al.*, 2013).

A related concept is nucleosome-mediated co-operativity or collaborative competition (Figure 1.5F). This model proposes that the combined binding energy of clustered transcription factors is sufficient to “out-compete” the stable interaction between histones and DNA. This co-operative binding enables chromatin to become accessible without the initial need for chromatin remodelling enzymes (Miller and Widom, 2003; Mirny, 2010).

Finally, the binding of one transcription factor may induce local DNA conformation changes, thereby enabling the binding of additional factors (Figure 1.5G). These changes may be subtle, for example involving small distortions of the DNA major groove (Kim *et al.*, 2013).

The complement of transcription factors that is expressed and/or is active in a cell type can influence which interactions occur. In the fruit fly *Drosophila melanogaster* (*Drosophila*), for example, the transcription factor protein mothers against dpp (Mad) becomes phosphorylated in response to cell signalling pathways. Phosphorylated Mad can then pair with cell-type-specific transcription factors on enhancers, for example Tinman in the dorsal mesoderm and Scalloped in the wing imaginal disc. In this way, a general signalling effector protein can impart specific responses by binding to DNA in combination with different partners that have restricted expression patterns (reviewed in Spitz and Furlong, 2012).

The complexity of transcription factor co-operativity is illustrated by several recent large-scale genome-wide studies. The ENCODE project, for example, has gathered ChIP-seq data across 5 human cell lines for 119 transcription-related factors (including 88 sequence-specific transcription factors, 16 factors associated with the core transcriptional machinery and 15 chromatin regulators) (Gerstein *et al.*, 2012). The study analysed co-associations between different factors, as well as changes in associations that occurred between promoter and distal regulatory regions. GATA1, for example, was found to co-associate with 6 primary partners (whose binding was consistently associated with GATA1 binding). In addition, several groups of local GATA1 partners were identified (associated with GATA1 only at specific subsets of GATA1 sites); these included the Jun proto-oncogene AP-1 transcription factor subunit (JUN) in one group and MYC-associated factor X (MAX) in another. Interestingly, GATA1 showed different co-association patterns in gene promoter and distal regulatory regions. The interaction with MAX, for example, was seen at promoter sites, along with

other factors such as Pol II, while the association with JUN (as well as JUND, JUNB and p300) was preferentially seen at distal sites. This indicates that different combinations of transcription factors can assemble at different genomic locations. These patterns strongly indicate that the co-binding of transcription factors is likely to be functionally relevant, although this is not directly demonstrated by the study.

As demonstrated by the above discussion, transcription factors can act co-operatively through a number of mechanisms. The same transcription factor may display multiple types of co-operativity. ETS1, for example, can co-operate through both homodimeric and heterodimeric binding and latent specificity. ER can co-operate through homodimeric binding, tethering and FOXA1-mediated chromatin remodelling. Co-operative mechanisms are likely to be specific to different cell types, as well as genomic locations, and the complexity of these co-operative interactions are only just beginning to be unravelled.

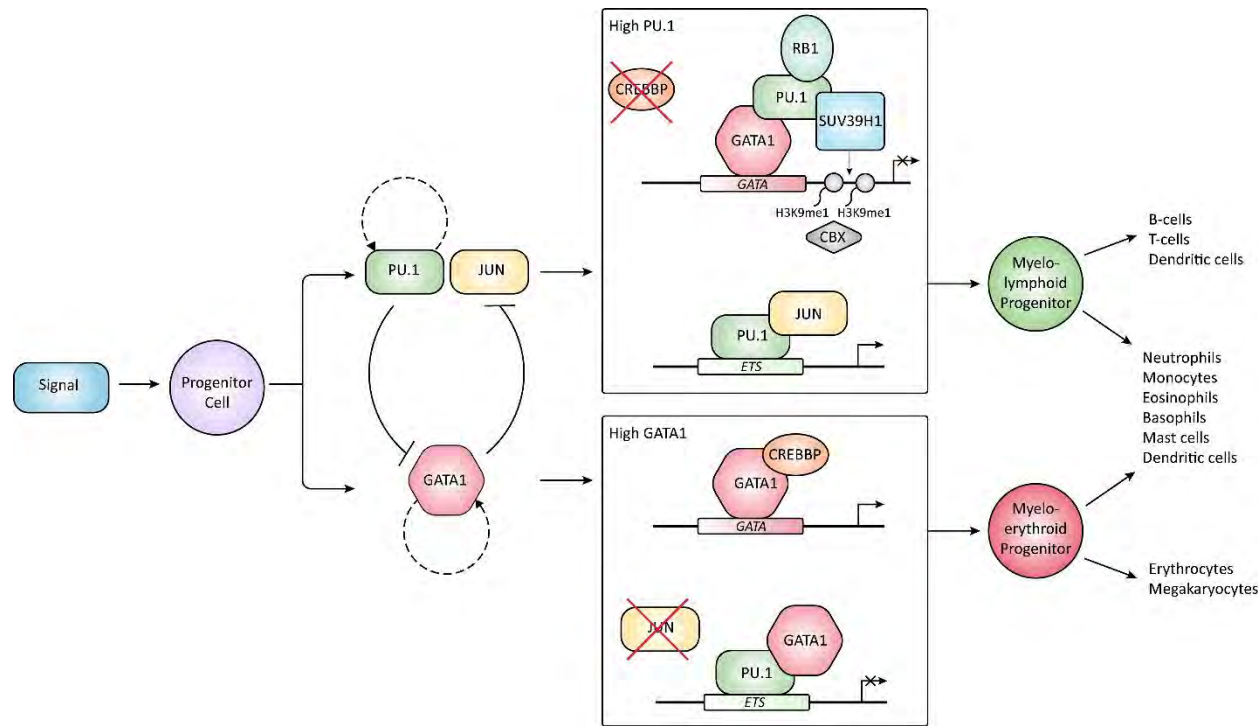
### **Transcription Factor Competition**

Some transcription factors act antagonistically rather than co-operatively. An example is the interaction between the myeloerythroid-specifying transcription factor GATA1 and the myelolymphoid-specifying ETS transcription factor Spi1 proto-oncogene (SPI1, also known as PU.1) (Arinobu *et al.*, 2007), shown in Figure 1.6. PU.1 and GATA1 have been shown to directly interact via their DNA-binding domains (Rekhtman *et al.*, 1999; Zhang *et al.*, 1999). An initiating event (for example, cytokine signalling) may alter the ratio of PU.1 and GATA1 in the progenitor cell. High GATA1 leads to GATA1-mediated inhibition of the interaction of PU.1 with its co-activator JUN, thereby suppressing the expression of myelolymphoid lineage genes (Zhang *et al.*, 1999). Conversely, PU.1 can interact with DNA-bound GATA1 to displace the co-activator CREB binding protein (CREBBP) and recruit proteins that create a repressive chromatin structure. These include RB transcriptional corepressor 1 (RB1) and suppressor of variegation 3-9 homolog 1 (SUV39H1), a histone methyltransferase that catalyses the repressive chromatin modification histone 3 lysine 9 monomethylation (H3K9me1). Chromatin remodellers such as chromobox 1 (CBX1) and 5 (CBX5) can subsequently bind to H3K9me1 to induce chromatin compaction (Stopka *et al.*, 2005). In this way, each transcription factor positively regulates the genes required for one lineage, while simultaneously repressing the expression of genes required for an alternative lineage. Both GATA1 and PU.1 also positively auto-regulate their own expression, further reinforcing the lineage pathway (Tenen, 2003).

A recent study has, however, cast doubt on this GATA1/Pu.1 model (Hoppe *et al.*, 2016). The study generated fluorescent reporter mice that expressed endogenous Gata1 tagged with mCherry and Pu.1 tagged with enhanced yellow fluorescent protein (eYFP). Based on flow cytometry and immunofluorescent analysis, the authors concluded that progenitor cells do not pass through a stage where Gata1 and Pu.1 protein are co-expressed (within the limits of detection of their methods). The haematopoietic stem cell population expresses intermediate Pu.1 but not Gata1, while the progenitor cell population in fact contains a mix of cell populations expressing either Gata1 or Pu.1 but rarely both. These progenitor cells (common myeloid progenitors) are already committed to their respective lineages. This indicates that Gata1 and Pu.1 cross-antagonism is not a primary mechanism of lineage specification. However, it is still possible that this antagonistic behaviour exists to reinforce cell lineage should aberrant expression of the opposing transcription factor arise.

Another example of transcription factor competition is the interaction of GATA binding protein 3 (GATA3) and forkhead box C1 (FOXC1) in breast cancer cells (Yu-Rice *et al.*, 2016). FOXC1 (commonly overexpressed in ER-negative breast cancer) negatively regulates ER expression, while GATA3 positively regulates ER expression. Overexpression of FOXC1 inhibits the binding of GATA3 to regulatory elements of the ER gene and this inhibition requires the forkhead (DNA-binding) domain of FOXC1. Several closely spaced GATA3 and FOXC1 binding sites have been identified in the enhancer and promoter regions of ER. In this way, high levels of FOXC1 can prevent the GATA3-initiated expression of ER in breast cancer. The mechanism of assisted loading does not seem to operate at this site and the reasons for this are currently unknown.

Finally, inhibition of one transcription factor by another may occur through tethering. This mechanism, known as transrepression, is used by glucocorticoid receptor (GR) to inhibit the activity of a number of DNA-bound immune-regulating transcription factors such as nuclear factor kappa B (NF $\kappa$ B) (reviewed in Ratman *et al.*, 2013).



**Figure 1.6: Model of GATA1 and PU.1 cross-antagonism during haematopoietic development**

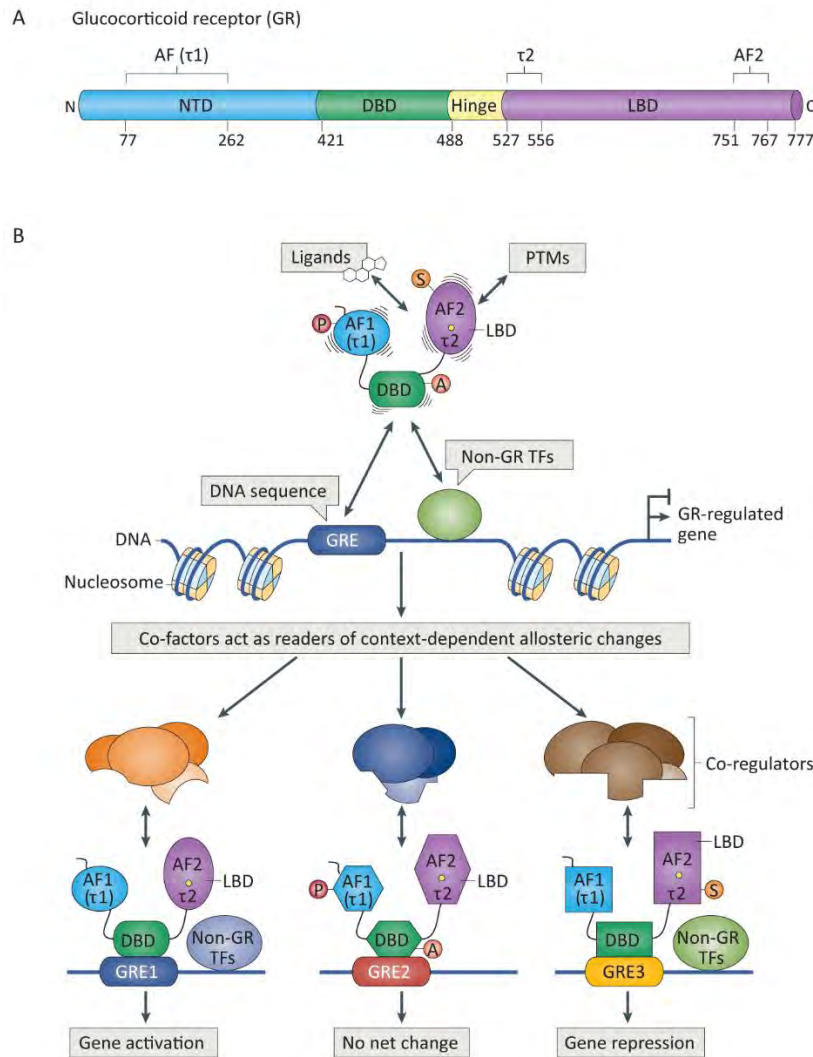
A signalling event initiates an alteration in the GATA1 / PU.1 ratio in the progenitor cell. PU.1 inhibits GATA1 activity, while GATA1 inhibits PU.1 activity. Each factor also auto-regulates their own expression (dashed arrows). In the top box (high PU.1), PU.1 binds to ETS recognition sequences, interacts with co-activator JUN and positively regulates myelolymphoid gene expression. PU.1 also represses myeloerythroid gene expression by binding to DNA-bound GATA1, displacing CREBBP and recruiting repressive co-factors. In the bottom box, GATA1 activates myeloerythroid gene expression by binding to GATA regulatory elements and also inhibits PU.1 activity by displacing JUN. GATA1, GATA binding protein 1; PU.1, Spi1 proto-oncogene; ETS, E-twenty-six; JUN, Jun proto-oncogene, AP-1 transcription factor subunit; CREBBP, CREB binding protein; RB1, RB transcriptional corepressor 1; SUV39H1, suppressor of variegation 3-9 homolog 1. (Arinobu *et al.*, 2007; Graf and Enver, 2009; Stopka *et al.*, 2005; Tenen, 2003; Zhang *et al.*, 1999).

## Interactions with Co-factors

Transcription factors may also act co-operatively with non-DNA-bound co-factors. These are proteins that contribute to activation and repression of gene transcription but do not have intrinsic DNA-binding activity (Young, 2011). Interactions with co-factors typically occur through activation or repression domains of sequence-specific transcription factors, which are distinct from the DNA-binding domain (Allen and Taatjes, 2015). The glucocorticoid receptor (GR), for example, contains three activation/repression regions (Figure 1.7A) (Weikum *et al.*, 2017). Transcription factors can generally interact with a wide range of co-factors. ER, as an example, has been shown to interact with over 100 co-factors (Manavathi *et al.*, 2014). Conversely, each co-factor may interact with many different sequence-specific transcription factors. Similar to co-operative activity with other transcription factors, these interactions provide an additional level of specificity for gene regulation.

Co-factors can have many functions and frequently assemble into multi-sub-unit complexes. Although they are a diverse group, co-factors can be classified based on their function. The glucocorticoid receptor (GR) co-factors, for example, can be divided into five main functional classes (Weikum *et al.*, 2017) and this classification is likely to be relevant to many types of transcription factors. The first functional class is structural and enzyme-interacting proteins. An example is the p160 family members steroid receptor co-activator (SRC) 1, 2 and 3 (reviewed in Weikum *et al.*, 2017). The SRC proteins have multiple protein-protein interaction domains and function as molecular scaffolds, promoting the formation of regulatory complexes. They are important co-regulators for a variety of nuclear receptors, including GR and ER. GR preferentially interacts with SRC2, which can contact GR at multiple interfaces. The N-terminal domain of SRC2, for example, can bind to the activation function domain 1 (AF1) regulatory region of GR. AF1 is an intrinsically disordered N-terminal domain and the binding of co-factors such as SRC2 leads to stabilisation. The central domain of SRC2 contains three LXXLL amino acid motifs (highly conserved protein-protein interaction motifs for interaction with nuclear receptors) and two transactivation domains. This region interacts with the activation function domain 2 (AF2) regulatory region of GR, which is contained within the C-terminal ligand-binding domain. The C-terminal domain of SRC2 can interact with additional co-factors, including histone acetyltransferases and histone methyltransferases.





**Figure 1.7: Co-factors as readers and effectors of context-dependent changes in glucocorticoid receptor conformation**

**Figure reproduced (and caption modified) with permission from Nature Publishing Group (Weikum *et al.*, 2017)**

(A) Structure of the 777 amino acid glucocorticoid receptor (GR), which consists of the amino-terminal domain (NTD), DNA-binding domain (DBD), hinge region and ligand-binding domain (LBD). Within these domains are segments that participate in transcriptional regulation: activation function domain 1 (AF1,  $\tau_1$ ), tau2 ( $\tau_2$ ) and activation function domain 2 (AF2). (B) A model for context-specific transcriptional regulation. Weikum *et al.* propose that GR serves as a scaffold that can adopt different conformations in response to regulatory inputs. These inputs are: Ligand binding and post-translational modifications (providing physiological context), DNA binding sequence (genomic context) and co-operating transcription factors (cellular context). The integrated actions of these inputs results in many possible GR conformations, subsequently determining which combination of co-factors can interact with GR. Lines around GR domains depict conformational changes imposed by signalling inputs. GRE, glucocorticoid response element; TF, transcription factor; A, acetylation; P, phosphorylation; S, sumoylation.

Another example of a co-factor in this structural class is the Mediator complex. Mediator interacts with a wide variety of transcription factors as well as with RNA Pol II. The complex forms a physical link between DNA-bound transcription factors and the core transcriptional machinery, allowing for the communication of regulatory signals from multiple pathways. Mediator has essential roles in transcription including the formation and stabilisation of the preinitiation complex and the regulation of promoter-proximal pausing, elongation and mRNA processing. Mediator also contributes to the formation of DNA loops between enhancers and promoters and the modulation of chromatin architecture through interactions with chromatin remodelling factors (Allen and Taatjes, 2015; Yin and Wang, 2014). In mammals, the Mediator complex is composed of 26 core sub-units. However, the sub-unit composition is variable and certain transcription factors are known to require specific Mediator sub-units in order to regulate transcription. The complex can also stably associate with a kinase module consisting of 4 sub-units, which may additionally regulate transcriptional processes through phosphorylation (Allen and Taatjes, 2015). It is clear from this brief discussion that Mediator has multiple and highly complex roles as a transcriptional co-factor.

The second functional class of transcriptional co-factors are the adenosine triphosphate (ATP)-dependent chromatin remodellers. These will be discussed in more detail in the following sub-section. Briefly, chromatin remodellers are recruited by transcription factors to modify the accessibility of DNA to other factors, such as the core transcriptional machinery and co-operating transcription factors. An example is the interaction of the SWI/SNF (SWItch / Sucrose Non Fermentable) family members SMARCA2 and BRG1 / SMARCA4 (SWI/SNF related matrix associated actin dependent regulator of chromatin subfamily a, members 2 and 4) with nuclear receptors including GR and ER. The activity of these chromatin remodellers is essential for the transcriptional activity of ER (reviewed in Green and Carroll, 2007; Weikum *et al.*, 2017).

The remaining functional classes of co-factors may be combined into a general group consisting of enzymatic histone modifiers. Once again, these will be discussed in further detail in another sub-section. Examples include histone acetyltransferases (HATs), histone deacetylases (HDACs), methyltransferases and demethylases. Many of the histone-modifying enzymes can also modify non-histone proteins to alter their activity (Weikum *et al.*, 2017). The effects of histone modifications include changes in chromatin structure and the recruitment of additional regulatory proteins that can recognise and interact with these modifications (Hassler *et al.*, 2016; Li *et al.*, 2007a).

There is also emerging evidence that DNA methyltransferases and enzymes that demethylate DNA may also act as transcriptional co-factors, forming an additional functional class. However, this remains an area of ongoing debate and will be discussed further in the relevant sub-section below.

Many co-factors function in diverse multi-sub-unit complexes. Complexes are often built around scaffold proteins (class 1 above), which facilitate the interaction between the transcription factor and other sub-units of the complex (Millard *et al.*, 2013). An example is the nuclear receptor co-repressor 2 (NCOR2, also known as SMRT), which forms transient and relatively weak interactions with nuclear receptors as well as other transcription factors. NCOR2 also forms stable, high-affinity interactions with three other core proteins, including HDAC3, forming the SMRT repressive complex (Watson *et al.*, 2012). Co-factor complexes may also contain proteins with recognition domains (for example, bromodomains which bind to acetylated lysines or chromodomains which bind to methylated lysines) in order to target the complex to specific chromatin locations. These domains may be within the enzymatic protein/s or in a different protein sub-unit of the complex (Millard *et al.*, 2013). The repressive Nucleosome Remodeling and histone Deacetylation (NuRD) complex, for example, contains amongst its members an HDAC sub-unit (HDAC1/2) and a chromatin remodelling sub-unit with a chromodomain (chromodomain helicase DNA binding protein (CHD) 3 or 4) (Torchy *et al.*, 2015). HDAC1 or 2 may also form part of the repressive LSD1-CoREST complex, in combination with the lysine demethylase 1A (KDM1A or LSD1), illustrating the interchangeable nature of co-factor complex sub-units (Meier and Brehm, 2014). In this way, transcription factors and their associated co-factors co-ordinate the assembly of the various enzymatic activities required for specific transcriptional regulation.

Co-factors may act as either activators or repressors depending on the genomic and cellular context. NCOR2 (SMRT), for example, was initially characterised as a co-repressor for unliganded retinoic acid receptor (RAR) and thyroid hormone receptor (TR) (see below) (Chen and Evans, 1995). In the breast cancer cell line MCF7, however, NCOR2 was shown to be required for transcriptional activation of specific genes by oestrogen receptor; this effect did not occur in the hepatocellular carcinoma cell line HepG2. The activating function of NCOR2 was independent of HDAC1 or HDAC3 expression, indicating that NCOR2 may act as a platform for the formation of alternative activating protein complexes in cell-specific contexts (Peterson *et al.*, 2007).

In a reciprocal manner, transcription factors can act as activators or repressors of transcription depending on the set of co-factors they associate with. A simple example

is the thyroid hormone receptor (TR). In the absence of ligand (thyroid hormone) binding, TR constitutively represses its target genes by binding to TR regulatory elements and interacting with repressive co-regulators such as NCOR2. When thyroid hormone binds to the receptor, the conformation of TR is altered in such a way that it can no longer bind to NCOR2. The repressive complex is released and TR interacts with activating co-regulators such as the histone acetyltransferase CREBBP (reviewed in Latchman, 2001). GR, which binds genes only upon hormone stimulation, can also activate certain genes and repress others in a cell-type-selective manner, an effect which is again dependent on the recruited co-factors.

An interesting model of GR transcriptional regulation, which enables this precise yet adaptable regulatory behaviour, is proposed by Weikum *et al.* (2017). This model, which could potentially apply to many types of transcription factors, places co-factors as the central “readers” of multiple regulatory events that modify GR conformation (known as allosteric regulation) (Figure 1.7B). These allosteric regulatory mechanisms (which will be discussed further in the section “Regulation of Transcription Factors”) include: (1) The binding of various ligands (for example, endogenous cortisol or the drug dexamethasone); (2) The specific DNA sequence to which GR is bound; (3) Post-translational modifications of GR (including phosphorylation, acetylation, sumoylation and others), and (4) Non-GR transcription factors (that may be cell-type-specific) that co-operate with GR by binding to composite regulatory elements or tethering GR to DNA (Figure 1.7B, top). The combination of these regulatory inputs shapes GR in different ways (for example, exposing different surfaces of the receptor), subsequently determining which sets of co-regulators are able to interact with GR (Figure 1.7B, bottom). In this way, a ubiquitously expressed transcription factor such as GR can receive information about the physiological context (ligand binding and signalling-mediated post-translational modifications), the genomic context (DNA sequence) and the cellular context (binding of other transcription factors) and integrate this information to produce a context-specific transcriptional response. The transcription factor itself, the authors argue, may simply be an adaptable scaffold upon which context-specific co-factors assemble to perform the regulatory tasks such as chromatin modification and recruitment of the transcriptional machinery. This provides an intriguing framework in which to consider the role of co-factors in transcriptional regulation.

The examples discussed above provide a glimpse of the diversity and complexity of the interactions of transcription factors with co-factors. Co-factors are essential to the function of transcription factors and, as proposed above, may even be the main

effectors of the context-specific information that is communicated by transcription factors. There are more than 350 co-factors that have been identified for the nuclear receptor family of transcription factors alone (Millard *et al.*, 2013) and this number continues to grow with the development of new technologies. GREB1, for example, was recently identified as a novel ER co-factor through a technique known as Rapid IMmunoprecipitation of Endogenous proteins (RIME) (Mohammed *et al.*, 2013). The function of GREB1 is currently unknown and this co-factor is just one example of many yet-to-be-identified co-regulators that may provide new insights into the mechanisms of transcriptional regulation.

### **Chromatin Remodelling**

An important mechanism of transcriptional regulation is the initiation of chromatin remodelling by transcription factors and associated co-factors. In eukaryotes, several metres of DNA fits into a nucleus that is only 2-10 micrometres in diameter. This is achieved by condensing DNA into repeating structures called nucleosomes, each consisting of 147 bp of DNA tightly wrapped around a core of 8 histone proteins (H3, H4, H2A and H2B in duplicate). Each nucleosome is joined by a 10-50 bp of linker DNA and collectively these repeating nucleosomal sub-units (about 30 million per human cell) are termed chromatin (Dechassa and Luger, 2011; Rothbart and Strahl, 2014). Linker histone H1 binds to the DNA as it enters and exits the core nucleosome and is important for stabilisation of higher-order chromatin structures (Hergeth and Schneider, 2015).

However, chromatin is not simply a static packaging mechanism for DNA; the structure of chromatin is highly dynamic and facilitates controlled access of transcriptional regulators to DNA (Voss and Hager, 2014). The interaction between transcription factors and chromatin is complex and reciprocal - transcription factors can both regulate and be regulated by chromatin structure. Chromatin represents a significant barrier to the binding of transcription factors, as the interaction between histone proteins and DNA is very stable (Li *et al.*, 2007a). Conversely, a number of transcription factors can initiate local reorganisation of the chromatin structure to overcome this barrier.

One model by which transcription factors may overcome this barrier is nucleosome-mediated co-operativity, also known as collaborative competition (Miller and Widom, 2003; Mirny, 2010). In this model, the binding of multiple transcription factors to closely-spaced regulatory elements (generally within one nucleosome) allows these factors to

effectively “out-compete” histones for the binding to DNA. This results in passive eviction of histones and increased chromatin accessibility. Additional co-factors, including chromatin remodelling enzymes, may subsequently be recruited by the DNA-bound transcription factors to stabilise the open state. An important (and yet unclear) aspect of this model is that it requires initial access of transcription factors to nucleosomal DNA, possibly through spontaneous DNA unwrapping and rewinding. In keeping with the dynamic nature of chromatin, nucleosomes have been shown to undergo spontaneous conformational shifts in which DNA unwinds from one end of the nucleosome. These changes occur on a rapid timescale, with DNA cycling between the fully wrapped state (for approximately 250 milliseconds) and the fully unwrapped state (for approximately 10-50 milliseconds) (Bucceri *et al.*, 2006). Nucleosomes can also undergo smaller conformational changes that affect DNA as it enters the nucleosome, known as “nucleosome breathing” (van Bakel, 2011). These transient events could provide a window of time in which sequence-specific transcription factors may access their DNA-binding sites and set in motion various chromatin remodelling events.

Another way in which transcription factors may regulate chromatin structure is through pioneer functions, which will be discussed further in a later sub-section. Briefly, pioneer transcription factors are able to bind to their target sites within nucleosomal DNA and therefore initiate regulatory events in previously silent chromatin (Zaret and Mango, 2016). Although the mechanisms are not completely understood, binding of pioneer factors results in increased chromatin accessibility for other transcription factors, thereby providing a mechanism for the initiation of cell-type-specific gene regulation. The ability of canonical pioneer factors such as FOXA1 to initiate local chromatin reorganisation does not require the activity of ATP-dependent chromatin remodellers, although these may subsequently be recruited along with additional co-factors (Zaret and Carroll, 2011).

One area of debate has been whether chromatin accessibility is actively regulated by transcription factor binding or whether it is pre-determined and merely permissive to transcription factor binding. Both situations are known to occur. Pioneer factors represent one end of the spectrum, while other transcription factors rely almost exclusively on pre-existing accessible chromatin for binding. An example of the latter is the glucocorticoid receptor (GR), which relies on pre-determined (although cell-specific) chromatin accessibility at 95% of hormone-induced binding sites (John *et al.*, 2011). Biggin (2011) argues that widespread and overlapping binding of transcription factors throughout the genome occurs due to locally permissive chromatin structure and that

many of these low-affinity binding sites (associated with low occupancy) are non-functional. However, other evidence points to the essential role of transcription factors, acting co-operatively, to promote accessible chromatin. The ENCODE project, for example, demonstrated a high degree of correlation between the combined binding signals of 42 transcription factors and the level of DNase hypersensitivity (an indicator of accessible chromatin) in K562 cells. This strong correlation implies that the binding of transcription factors is what is driving the accessible state (Thurman *et al.*, 2012).

Once bound to DNA, transcription factors can recruit co-factors such as chromatin remodellers and histone modifiers to alter the chromatin landscape. Chromatin remodellers can hydrolyse adenosine triphosphate (ATP) to reposition nucleosomes along the DNA in a process known as nucleosome sliding. In addition, some remodellers can disassemble nucleosomes, exchange core histones for histone variants or assemble nucleosomes *de novo*, although these processes are all likely to be underpinned by the sliding activity (Mueller-Planitz *et al.*, 2013). The basic sliding mechanism is driven by the catalytic (ATPase) domain, which is closely related to DNA and RNA helicases. Additional domains anchor the remodeller to the nucleosome surface and “ratchet” the DNA one base pair at a time. This process is driven by conformational changes in the remodeller related to ATP hydrolysis (Mueller-Planitz *et al.*, 2013).

Chromatin remodelling proteins also contain accessory domains, which can be used to classify remodellers into four main structural families. These are the SWItch / sucrose non fermentable (SWI/SNF), imitation switch (ISWI/SNF2L), chromodomain (CHD) and inositol (INO80) families. Different families are associated with different general functions. The SWI/SNF family, for example, is often associated with transcriptional activation, whereas the ISWI/SNF2L family is associated with transcriptional repression (Koster *et al.*, 2015). The accessory domains of chromatin remodellers are essential to the function of remodellers. The ISWI/SNF2L and CHD families, for example, contain C-terminal SANT and SLIDE domains, which bind to DNA and assist with nucleosome anchoring (Mueller-Planitz *et al.*, 2013). The additional domains also help to fine-tune the function of chromatin remodellers by linking their binding to particular histone and/or DNA modifications. The SWI/SNF family members, including SMARCA2 and SMARCA4, contain a bromodomain, which can interact with acetylated lysines, while the CHD family members contain two chromodomains, which can bind to methylated histones or DNA (Koster *et al.*, 2015). The frequent incorporation of chromatin remodellers into complexes, containing additional sub-units with varying functions, also

assists with the specific targeting of remodelling function.

The targeting of chromatin remodellers to DNA through recognition of specific histone modifications is just one example of the extensive interconnections between transcription factors, chromatin structure, histone modifications and DNA modifications. Chromatin structure, for example, may be also be affected indirectly by transcription factor recruitment of histone acetyltransferases; the acetylation of histones neutralises their positive charge and results in a loosening of the interactions with negatively-charged DNA (van Bakel, 2011). The deposition of strongly repressive histone modifications such as H3K9me2/3 results in a highly compressed chromatin structure that is inaccessible to transcription factors, including many pioneer factors (Iwafuchi-Doi and Zaret, 2014). Methylated DNA may recruit histone deacetylases and chromatin remodelling complexes through methyl-DNA binding proteins to decrease chromatin accessibility (Hassler *et al.*, 2016). The complex interactions between these various processes facilitate specific and precise regulation of gene expression by transcription factors.

## **Histone Modifications**

Histone modifications are intimately related to chromatin structure. The nucleosome core is composed of a histone octamer, containing two dimers of H2A and H2B and a tetramer of H3 and H4. Each histone protein contains a globular domain as well as a flexible, apparently unstructured N-terminal tail that protrudes from the nucleosome (Nightingale, 2011). Many post-translational modifications (PTMs) have been identified on the easily accessible histone tails (particularly H3 and H4), however increasing evidence suggests that modifications also occur in the globular domains. There are at least 12 types of modifications at 130 different locations so far identified, affecting residues on all four core histones as well as linker histone H1 and histone variants (Stricker *et al.*, 2017). Modifications include methylation, acetylation, phosphorylation, ubiquitylation, sumoylation, formylation, oxidation, crotonylation, hydroxylation, butyrylation, propionylation, ADP-ribosylation, proline isomerisation and citrullination. They are primarily located on lysine or arginine residues but may also be found on tyrosine, serine or glutamine residues (Rothbart and Strahl, 2014; Stricker *et al.*, 2017; Tessarz and Kouzarides, 2014). The functional associations of many of these modifications is an ongoing area of research.

Histone modifications are catalysed by enzymes such as histone acetyltransferases and methyltransferases. In many cases, corresponding enzymes that remove these



modifications have also been identified. Histone methylation was previously considered to be stable and irreversible, however its dynamic nature was revealed with the relatively recent discovery of histone demethylases, such as lysine demethylase 1A (KDM1A, also known as LSD1) (Shi *et al.*, 2004) and the JumonjiC domain-containing demethylases (Tsukada *et al.*, 2006). It is now recognised that histone modifications, just like transcription factor binding events, are often highly dynamic and represent the balance of ongoing enzyme activity at a given location. Histone acetylation, for example, has a half-life of only a few minutes (Nightingale, 2011). Histone acetyltransferases (HATs) and deacetylases (HDACs) frequently co-localise at active genes, where HDACs function to “reset” chromatin for future cycles of transcription (Wang *et al.*, 2009b).

Histone modifying enzymes may be recruited to regulatory elements by various mechanisms. Firstly, and most relevant to this discussion, DNA-bound sequence-specific transcription factors may recruit histone modifiers as co-factors. FOXA1, for example, recruits lysine methyltransferase 2C (KTM2C, also known as MLL3) to enhancers, where it catalyses H3K4me1 (Jozwik *et al.*, 2016). Another example is the interaction of snail family transcriptional repressor 1 (SNAI1) with the lysine demethylase KDM1A. The N-terminal SNAG domain of SNAI1 bears a strong resemblance to the N-terminus of histone H3, acting as a molecular “hook” to recruit KDM1A to its binding sites (Lin *et al.*, 2010). A second mechanism of recruitment of histone modifying enzymes is through the recognition of DNA features, such as the CpG content and methylation status. Proteins containing the zinc finger-CxxC (ZF-CxxC) domain specifically bind to non-methylated CpG islands, commonly found in vertebrate promoters. This domain is found in several histone modifying proteins, including the lysine methyltransferases MLL1 and MLL2, the lysine demethylases KDM2A and KDM2B and the CXXC1 sub-unit of the SET domain containing 1 (SETD1) methyltransferase complex (Long *et al.*, 2013). Thirdly, the core transcriptional machinery may recruit histone modifying enzymes, coupling specific modifications to active transcription. An example is the yeast methyltransferase Set2, which binds to the phosphorylated CTD of RNA Pol II and catalyses H3K36 methylation in actively transcribed gene bodies (Krogan *et al.*, 2003). Finally, a number of histone modifying complexes can recognise and bind to specific histone modifications. Recognition of their own modification results in the establishment of positive feedback loops. The SETD1 sub-unit CXXC1, for example, contains a plant homeodomain (PHD) finger domain (in addition to the ZF-CxxC domain), which recognises the H3K4me3 modification catalysed by the SETD1 complex (Zhang *et al.*, 2015). Another example is

the Polycomb repressive complex 2, which catalyses the repressive histone modification H3K27me3 and can also bind to this modification, contributing to the formation of repressive Polycomb domains (Blackledge *et al.*, 2015).

The previous paragraph introduces a number of concepts related to histone modifications. The first of these is that, just like for chromatin structure, there is a two-way interaction between transcription factors and histone modifications. Transcription factors can regulate modifications through the recruitment of histone modifiers but can also be regulated by these modifications, through direct and indirect effects on chromatin structure. A similar relationship exists for the core transcriptional machinery, which can recruit histone modifiers through the Pol II CTD but may also be recruited by direct binding of TFIIID sub-units to the H3K4me3 modification (through PHD domains) or to acetylated residues (via bromodomains) (Nightingale, 2011; Vermeulen *et al.*, 2007). A second concept is that many chromatin-interacting proteins contain “reader” domains, such as PHD and bromo-domains, that recognise specific modifications (Table 1.1). This may be an important mechanism by which histone modifications lead to functional outcomes. In addition, many histone-modifying enzymes (for example SETD1A/B and Polycomb proteins) act in multi-sub-unit complexes, similar to chromatin remodellers, flexibly bringing together multiple recognition domains and enzymatic activities. It is evident that there is a complex interaction between the core transcriptional machinery, transcription factors, histone modifications, chromatin structure and DNA modifications.

The advent of technologies such as ChIP-seq has revealed that specific histone modifications are tightly correlated with functional genomic elements (Table 1.1). Promoters, for example, are enriched for H3K4me3, while active regulatory elements are characterised by H3K9 and H3K27 acetylation and repressed chromatin by the mutually exclusive methylation of H3K9 or H3K27 (Nightingale, 2011). This has greatly assisted in the identification of regulatory elements such as enhancers (Zhou *et al.*, 2011). However, whether these modifications actively direct chromatin states, or instead reflect processes such as transcription that influence chromatin state, is an ongoing area of controversy.

Many early studies focused on the direct structural effects of histone modifications on nucleosome structure. Histone tail acetylation, for example, causes a neutralisation of lysine and arginine positive charges, resulting in a loosening of histone interactions with negatively-charged DNA and an increase in transcription factor activity (Lee *et al.*, 1993; Oliva *et al.*, 1990; Rice and Allis, 2001; Vettese-Dadey *et al.*, 1996). The

identification of modified residues within the globular domains, particularly on the lateral surface where histone proteins make many contacts with DNA, has led to a recent revival of research into these direct structural effects (Tessarz and Kouzarides, 2014). This will be discussed in further detail in the histone modifications sub-section of “Regulation of Transcription factors” (Part 1.2).

In 2000, a new functional perspective emerged, focusing on the potential instructive capacity of histone modifications. In the “histone code” hypothesis, Strahl and Allis (2000) propose that combinations of histone modifications are instructive and are interpreted by effector proteins (“readers”) to produce functional outcomes. The hypothesis was supported by the discovery of acetyl-binding by bromodomains just one year earlier (Dhalluin *et al.*, 1999). The histone code provided an explanation (in the form of combinatorial recognition) for how the same modification might be associated with opposing functions, for example the association of histone H3 serine 10 phosphorylation (H3S10p) with both chromatin condensation in mitosis and decompaction in transcription (Strahl and Allis, 2000).

A basic premise of the histone code hypothesis is the recognition of modifications (and their combinations) by effector proteins containing specific binding domains. Soon after the hypothesis was initially proposed, a number of methyl-binding domains were discovered, including PHD, MBT, chromo-, tudor and WD40 domains (Table 1.1). Many recognition domains are highly specific not only for the type of modification but also the degree (for example, mono-, di- or tri-methylation) (Nightingale, 2011). Recognition domains are commonly found in histone modifying and chromatin remodelling complexes. The SWI/SNF chromatin remodellers SMARCA2 and SMARCA4, for example, each contain a bromodomain, while the nucleosome-remodelling factor (NURF) complex sub-unit BPTF contains a bromodomain and two PHD domains (Koster *et al.*, 2015; Nightingale, 2011). The BPTF bromodomain binds to acetylated residues on histone H4, while the second BPTF PHD finger mediates specific binding to trimethyl-lysine (H3K4me3) (Li *et al.*, 2007b). *In vitro* experiments demonstrate that H3K4me3 and H4K16 acetylation (H4K16ac) within the same nucleosome provide the optimal template for BPTF binding (Ruthenburg *et al.*, 2011). BPTF is just one example of a “reader” protein containing multiple recognition domains, which may impose the combinatorial specificity underpinning the histone code hypothesis.

However, others argue that histone modifications are a consequence of processes that act on chromatin, such as transcription, rather than their driving force (Henikoff and Shilatifard, 2011; Rando, 2012). It is argued, for example, that recognition domains

bind to modified histones with very low affinity, casting doubt on the potential of these modifications to actively recruit effector proteins. In addition, combinatorial increases in affinity for proteins containing multiple recognition domains are often minimal. BPTF, for example, binds to doubly modified nucleosomes (H4K16ac and H3K4me3) with approximately two-fold increased affinity compared to H3K4me3 alone, implying that discrimination between these two states by BPTF *in vivo* is likely to be low (Rando, 2012; Ruthenburg *et al.*, 2011). Furthermore, the complexity of possible combinations of histone modifications does not appear to be utilised *in vivo*. This can be seen in ChIP-seq data sets profiling multiple modifications, which can be compressed into a small number of chromatin “states” with minimal loss of information (reviewed in Rando, 2012). Finally, manipulations of histone modifications (for example, through the deletion of histone-modifying enzymes or mutations in histone proteins) have produced some intriguing results. Deletion of the Polycomb repressive complex 2 (PRC2, which catalyses H3K27me3) in mouse embryonic stem cells, for example, causes only minor changes in gene expression (Riising *et al.*, 2014). Conditional deletions in other cell types have also demonstrated surprisingly specific (rather than broadly dysregulated) transcriptional consequences and phenotypes (reviewed in Rando, 2012; Stricker *et al.*, 2017). These results may imply the absence of a causative code or that only a small number of modifications have direct causative effects, which are dependent on the cellular context (Stricker *et al.*, 2017). An alternative model to the histone code is that histone modifying enzymes are recruited by factors such as sequence-specific transcription factors and the core transcriptional machinery, reflecting and facilitating processes such as transcription but not actively directing them. Modifications, rather than actively recruiting effector proteins, may then act to stabilise the binding of additional proteins or even activate their enzymatic activity through allosteric effects (Henikoff and Shilatifard, 2011; Rando, 2012).

The current understanding of histone modifications is far from complete. Superimposed on this already busy landscape are further complexities, including novel types of modifications, globular domain modifications, asymmetrical combinations of modifications (affecting adjacent histone tails within the same nucleosome), incorporation of histone variants, and cross-talk with other histone and DNA modifications (Rothbart and Strahl, 2014). New targeted epigenome editing techniques, for example based on the CRISPR-Cas system, may help to more precisely define the causal functions of histone modifications (Stricker *et al.*, 2017).

**Table 1.1: Examples of common post-translational histone modifications**

Modification	Enzymes* ("writers")	Genomic location	Functional association	Binding domains	Removers* ("erasers")
METHYLATION					
H3K4me1	SETD1A, SETD1B, KMT2A, KMT2B, KMT2C,	Enhancers and downstream of TSS	Gene activation Poised state	Plant homeodomain (PHD) Malignant brain tumour (MBT) domain Tudor domain Chromodomain WD40 domain Ankyrin repeats	KDM1A, KDM1B, KDM2B, KDM5A, KDM5B, KDM5C, KDM5D, RIOX1
H3K4me2	KMT2D, KMT2E, SMYD1, SMYD2,	Promoters and enhancers	Gene activation Poised state		
H3K4me3	SMYD3, PRDM7, PRDM9, SETD7	Promoters and TSS	Gene activation		
H3K36me3	SETD2, ASH1L, NSD1	Transcribed genes (3' enrichment)	Gene activation Splicing		KDM4A, KDM4B, KDM4C, KDM4D, RIOX1
H3K79 methylation	DOT1L	Transcribed genes (5' enrichment)	Gene activation		Unknown
H3K9 methylation	SETDB1, SETDB2, EHMT1, EHMT2, SUV39H1, SUV39H2, PRDM2	Inactive promoters, repressed chromatin	Gene repression		KDM1A, KDM3A, KDM3B, KDM4A, KDM4B, KDM4C, KDM4D, KDM7A, PHF2, PHF8, RIOX2
H3K27me1	EZH1 and EZH2 (in PRC2), NSD2 and NSD3 (possibly), EHMT1, EHMT2	Inactive promoters and enhancers, repressed chromatin	Gene repression		KDM6A, KDM6B, KDM7A, PHF2, PHF8
H3K27me2/3		Repressed chromatin			
H4K20me1	KMT5A	5' end of genes	Gene activation or repression (unclear)		PHF8
H4K20me2/3	KMT5B, KMT5C	Repressed chromatin	Gene repression		Unknown
ACETYLATION					
H3K9ac	EP300, CREBBP, KAT2A, KAT2B, KAT6A	Active regulatory elements (particularly promoters)	Gene activation	Bromodomain Selected plant homeodomains (PHD)	Most histone deacetylases with the exception of SIRT6 (H3K9 specific) and SIRT7
H3K27ac		Active regulatory elements (enhancers and promoters)	Gene activation		

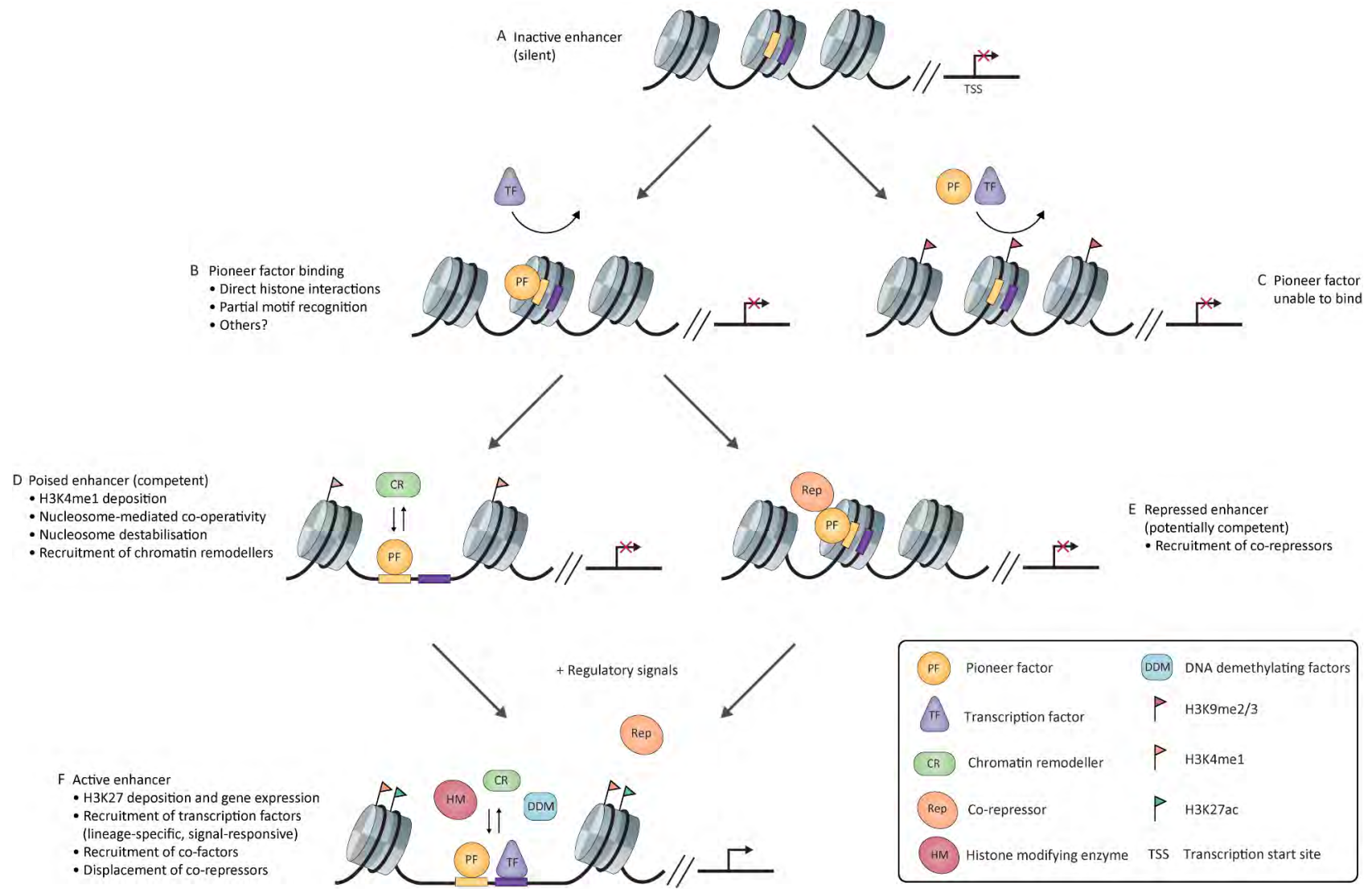
References: (Cheng, 2014; Dunham *et al.*, 2012; Fog *et al.*, 2012; Green and Carroll, 2007; Herz *et al.*, 2013; Houtkooper *et al.*, 2012; Kooistra and Helin, 2012; Liu *et al.*, 2016; Marmorstein and Zhou, 2014; Miller and Grant, 2013; Mozzetta *et al.*, 2015; NCBI Resource Coordinators, 2016; Nightingale, 2011; Seto and Yoshida, 2014; The Uniprot Consortium, 2017; Volkel and Angrand, 2007; Weikum *et al.*, 2017; Wu *et al.*, 2013; Yang and Bedford, 2013; Zhou *et al.*, 2011).

## Pioneer Functions

Pioneer transcription factors have the ability to recognise and bind to their target sequences in nucleosomal DNA, initiating a local increase in chromatin accessibility and facilitating the binding of additional regulatory factors. In this way, pioneer factors can initiate regulatory events in previously silent chromatin, for example during normal development, cell fate reprogramming and steroid hormone stimulation (Iwafuchi-Doi and Zaret, 2014; Zaret *et al.*, 2016; Zaret and Mango, 2016) (Figure 1.8).

An essential feature of pioneer factors is the ability to interact with closed chromatin. Within nucleosomes, one face of the DNA helix long axis is occluded by contacts with core histones. Forkhead box A (FOXA) proteins were one of the first family of pioneer factors to be described and have a winged-helix DNA-binding domain that resembles the globular domain of linker histone H1 (Cirillo *et al.*, 1998). The structure of this domain allows FOXA proteins to interact with the single exposed face of the DNA helix in a similar way to linker histones. This interaction is further stabilised by the FOXA C-terminal domain, which binds to histones H3 and H4 (Cirillo *et al.*, 2002).

Another mechanism by which pioneer factors interact with closed chromatin is through the recognition of partial DNA motifs exposed on the nucleosome surface (Soufi *et al.*, 2015). The pioneer factor POU class 5 homeobox 1 (POU5F1, also known as OCT4) binds to DNA through a bipartite POU domain, consisting of two sub-domains connected by a linker region. At nucleosome-depleted (accessible) sites, each POU sub-domain binds to one half of a canonical 8 bp OCT4 motif. However, at sites of high nucleosome occupancy (inaccessible chromatin), OCT4 recognises a 6 bp motif that resembles one half of the canonical motif. OCT4 can therefore initiate binding at these sites using a single POU sub-domain, which requires interaction with only one face of the DNA helix; this leaves the opposite face of the helix potentially free to interact with the histone octamer. Similar mechanisms were identified for other induced pluripotent stem cell factors, including SOX2 and KLF4. Therefore, the ability of a transcription factor to act as a pioneer factor may be related to its ability to adapt its binding to a reduced motif that is compatible with nucleosome structure. This mechanism of pioneering is different to that of FOXA1, which binds to a similar motif regardless of baseline chromatin accessibility (Soufi *et al.*, 2015).



**Figure 1.8: Pioneer factors interact with nucleosomal DNA and “bookmark” enhancers for future activation**

*(previous page)*

(A) Inactive enhancers contain nucleosomes that lack active histone modifications. There is no transcription of regulated genes. (B) Pioneer factors are able to bind to their motifs in DNA that is inaccessible to other transcription factors through various mechanisms (listed). (C) Binding of pioneer factors can be inhibited by certain chromatin features. An example is histone H3 lysine 9 dimethylation or trimethylation (H3K9me2/3), associated with repressive chromatin. (D) Once bound to DNA, pioneer factors can initiate a local increase in chromatin accessibility. Histone H3 lysine 4 monomethylation (H3K4me1) is a feature of poised enhancers, which are primed for future activation in response to regulatory signals. (E) Some pioneer factors may recruit co-repressors to guard against premature expression of lineage-specific genes. (F) In response to regulatory signals, additional factors are recruited to enhancers, including lineage-specific transcription factors and co-factors. Co-repressors are displaced. Transcription of regulated genes is activated, accompanied by acetylation of histone 3 lysine 27 (H3K27ac).

Pioneer factors have also been proposed to have longer residence times than other transcription factors on DNA. Green fluorescent protein (GFP)-tagged FOXA1 and FOXA2 have been shown by fluorescence recovery after photobleaching (FRAP) to move more slowly than other transcription factors in the nucleus. This suggests that pioneer factors may have more stable interactions with nucleosomes, enhancing their ability to scan closed chromatin for binding sites (Sekiya *et al.*, 2009). However, a recent study using a single-molecule tracking (SMT) has questioned this property of pioneer factors (Swinstead *et al.*, 2016a). Using SMT, which is argued to be more precise and direct than FRAP (Zaret *et al.*, 2016), this study found that the residence time of FOXA1 on chromatin is comparable to that of ER and GR, with a “slow” component residence time of 10.8 seconds (compared to 8.4 and 8.8 seconds for ER and GR respectively) (Swinstead *et al.*, 2016a). Therefore, it appears unlikely that the binding dynamics of FOXA1 contribute to its pioneering ability.

Once bound to DNA, pioneer factors initiate an increase (or in some cases a decrease) in chromatin accessibility, influencing the binding of additional factors. There are several mechanisms by which this may occur. Firstly, pioneer factors may enhance the binding of other factors through nucleosome-mediated co-operativity (Zaret and Carroll, 2011). Pioneer factors are more easily able to overcome the initial barrier posed by chromatin, leading to a reduction in the energy required for additional factors to overcome the histone-DNA interaction. Secondly, pioneer factors may directly interact with histones to destabilise nucleosome structure. FOXA1 has been shown to directly



bind to nucleosomes *in vitro* and to increase chromatin accessibility without the requirement for ATP-dependent chromatin remodellers (Cirillo *et al.*, 2002). Intriguingly, a recent study has demonstrated that liver-specific enhancers (even when active) are more likely than ubiquitous enhancers to contain nucleosomes. However, the binding of FOXA1 to these enhancers creates an “accessible” nucleosome structure, which can still be bound by liver-specific transcription factors, through the displacement of linker histone H1 (Iwafuchi-Doi *et al.*, 2016). Finally, the recruitment of ATP-dependent chromatin remodellers is essential for the pioneering activity of a subset of factors, including ER, GR and GATA3 (Swinstead *et al.*, 2016b). Classic pioneer factors such as FOXA1 may also recruit chromatin remodellers to stabilise the open state.

Pioneer factors can also decrease chromatin accessibility and repress gene transcription. An example is the interaction of FOXA1 with transducin like enhancer of split 3 (TLE3, also known as GRG3). The recruitment of TLE3 by FOXA1 to chromatin results in compaction of three to four nucleosomes and decreases the expression of genes regulated by FOXA1 (Sekiya and Zaret, 2007). FOXA factors are essential for liver development during embryogenesis, while TLE3 is expressed in undifferentiated embryonic endoderm cells and is downregulated during hepatic differentiation. It has been proposed that recruitment of the TLE3 co-repressor is important for maintaining liver-specific gene silencing (while retaining competence for future expression), therefore guarding against premature expression of hepatic genes. The subsequent loss of TLE3 contributes to the ability of FOXA1 to initiate the liver-specific gene expression program (reviewed in Zaret and Carroll, 2011).

Pioneer factors are essential to the normal development of many tissues. FOXA and GATA factors bind to the liver-specific albumin (*Alb1*) enhancer during mouse liver development and were two of the first pioneer factors to be described. Importantly, FOXA and GATA factors bind to the *Alb1* enhancer in undifferentiated gut endoderm prior to the transcriptional activation of hepatic genes (reviewed in Zaret and Carroll, 2011). This association with chromatin prior to activation is proposed to be an additional characteristic of pioneer factors, priming cell-specific enhancers for future activation in response to regulatory signals. Recently, the role of FOXA factors in endodermal development has been further studied using an *in vitro* differentiation timecourse in human embryonic stem cells (hESCs) (Wang *et al.*, 2015a). Histone modifications and transcription factor binding were mapped as the cells progressed from hESCs through four stages of pancreatic development (definitive endoderm (DE), primitive gut tube (GT), posterior foregut (FG) and pancreatic endoderm (PE)). Cells

were also differentiated along other endodermal lineages, including lung and liver. Histone modifications were used to classify enhancers at various stages as poised (H3K4me1 modification only) or active (H3K4me1 and H3K27ac modifications). Many enhancers became poised (that is, acquired H3K4me1) during the transition from DE to GT, however a significant number of these never became active and were shown to have a role in the differentiation of more mature pancreatic cells (such as islets), lung or liver. Enhancers that became active during the pancreas-specific stages (FG and PE) showed enrichment for FOXA1 motifs, with recruitment of FOXA1/2 to these sites at the GT stage prior to enhancer activation. FOXA1/2 were also recruited to liver-specific enhancers prior to enhancer activation when cells were differentiated along the hepatic lineage. Based on these findings, the study concluded that the ability of cells to initiate cell identity gene expression in response to regulatory inputs (developmental competence) is established through the poised chromatin state (characterised by H3K4me1). The pioneer factor FOXA1 interacts with H3K4me1 and “bookmarks” cell identity genes, establishing competence for a number of different possible endodermal lineages. In response to tissue-specific signals (for example, pancreatic growth factors), lineage-specific transcription factors such as pancreatic and duodenal homeobox 1 (PDX1) are recruited to these bookmarked sites, activating the enhancer and initiating tissue-specific gene expression. This provides a mechanism by which cells can correctly and efficiently respond to extrinsic developmental signals. Importantly, this model also emphasises the hierarchical yet co-operative nature of transcription factor function. Pioneer factors are not necessarily cell-type-specific but the sequential binding of additional signal-responsive factors enables combined regulation of a defined set of lineage-specific genes.

The action of pioneer factors is associated with a number of histone and DNA modifications, although it is debated as to whether these modifications guide pioneer factors to their binding sites or arise as a consequence of pioneer factor binding. One example is the histone modification H3K4me1/2 which, as discussed above, is associated with poised enhancers. Several studies suggest that FOXA1 recognises the H3K4me1/2 modification and that this modification is required for binding (Lupien *et al.*, 2008; Wang *et al.*, 2015a). In this scenario, it is unclear how the cell-type-specific H3K4me1/2 pattern is initially established. Other studies, however, have demonstrated that FOXA1 can facilitate the deposition of H3K4me1/2. Ectopic expression of FOXA1 in the ER- and FOXA1-negative breast cancer cell line MDA-MB-231, for example, results in the establishment of H3K4me1/2 at FOXA1 binding sites (Serandour *et al.*, 2011). In the MCF7 breast cancer cell line, endogenous FOXA1 recruits lysine

methyltransferase 2C (KMT2C, also known as MLL3), which catalyses H3K4me1/2 at enhancers. MLL3 silencing in MCF7 cells results in decreased enhancer H3K4me1/2 and compromised ER-induced gene regulation (Jozwik *et al.*, 2016). In this second scenario, however, it is unclear how FOXA1 binding is restricted to just a fraction of potential genomic binding sites to establish cell-type-specific H3K4me1/2 patterns. It is also possible that both of these mechanisms operate simultaneously to stabilise and reinforce FOXA1 binding.

DNA methylation poses a similar conundrum. FOXA1 binding is enhanced by low levels of DNA methylation (Serandour *et al.*, 2011). In addition, FOXA1 has been shown to be actively involved in DNA demethylation through the recruitment of DNA repair proteins that contribute to this process (Zhang *et al.*, 2016c). The pioneering potential of another transcription factor, nuclear respiratory factor 1 (NRF1), is inhibited by DNA methylation and it is proposed that the ability to demethylate DNA may be an important characteristic of true pioneer factors (Domcke *et al.*, 2015; Zhu *et al.*, 2016). It is evident that the order of events and the additional regulators involved in enhancer histone and DNA modifications are yet to be established.

In summary, pioneer factors have the ability to interact with nucleosomal DNA, modify chromatin accessibility (through intrinsic or co-operative mechanisms), and specify cell lineage. “Bookmarking” of enhancers by pioneer factors establishes transcriptional competence so that genes may be effectively and rapidly activated in response to specific regulatory signals during development or hormone stimulation. The mechanism by which this occurs is currently unclear, as recent research suggests that pioneer factors do not stably interact with chromatin; other possible mechanisms of “bookmarking” include modifications of histones and DNA by recruited co-factors. It has been proposed that the recruitment of ATP-dependent chromatin remodellers can enable many transcription factors to function as pioneer factors, through the mechanism of assisted loading. However, it is also evident that transcription factors vary widely in how effectively they can target nucleosomal DNA and subsequently recruit additional factors (Soufi *et al.*, 2015; Zaret *et al.*, 2016). It seems likely that chromatin remodellers are important in pioneer function although the extent to which this mechanism operates is currently unknown. Elucidating the mechanisms by which different types of transcription factors initiate cell-type-specific gene transcription is an exciting area of ongoing research.

## DNA Methylation

DNA, just like histones, may be covalently modified, superimposing an additional level of regulation on gene expression. DNA methylation is the addition of a methyl group to carbon 5 of cytosine to produce 5-methylcytosine (5mC). It is catalysed by the DNA methyltransferases DNMT1, DNMT3A and DNMT3B, along with the catalytically inactive co-regulator DNMT3L (Frauer *et al.*, 2011). Demethylation is a more difficult process and the known mechanisms, only recently identified, involve excision of the modified nucleotide rather than direct enzymatic removal of the methyl group. The ten-eleven translocation (TET) enzymes, for example, convert 5mC to 5-hydroxymethylcytosine (5hmC) and subsequently to 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC) (Ito *et al.*, 2010; Tahiliani *et al.*, 2009). 5fC and 5caC can then be removed by enzymatic base excision (mediated by DNA glycosylases) followed by DNA repair, restoring an unmethylated template (He *et al.*, 2011). Passive demethylation can also occur by preventing the methylation of DNA after replication (Frauer *et al.*, 2011).

Most methylation in humans occurs at CG (CpG) dinucleotides (Hassler *et al.*, 2016). Methylation in other contexts (for example, CHG or CHH, where H is any nucleotide other than G) has also been observed, and is widespread in fungi and plants, however the functional relevance in humans is unknown (Dunham *et al.*, 2012; Jones, 2012). Overall, the human genome contains significantly fewer CpG dinucleotides than would be expected and this has been proposed to result from the spontaneous deamination of methylated cytosine to thymine (Bird, 1980). Despite this relative depletion, there are 30 million CpGs present at sites throughout the genome, including promoters, intergenic sequences, gene bodies and repetitive elements (Chatterjee and Vinson, 2012; Frauer *et al.*, 2011). The occurrence of CpGs falls into two main patterns - CpG islands (CGIs) and non-CpG islands. CGIs contain dense clusters of CpG dinucleotides and mostly occur at gene promoters, with about 60% of all human genes containing a CGI (Frauer *et al.*, 2011). The majority of CGIs are never methylated and remain unmethylated in somatic cells (Jones, 2012). Conversely, CpGs outside of CGIs are present at a variety of genomic sites and are often highly (although dynamically) methylated. The most variable sites of methylation in the human genome are located in low-CpG intergenic regions (including enhancers) and gene bodies (Dunham *et al.*, 2012).

Much of the previous research into DNA methylation has focused on CGIs in gene promoters and this has shaped the general perception that DNA methylation leads to

gene repression. Emerging evidence, however, suggests that DNA methylation can be associated with a variety of effects, dependent on genomic location, CpG density and the interaction of various “reader” proteins that can specifically recognise the methylation state of DNA. Methylation can even be associated in some cases with gene activation (Jones, 2012). In addition, the dynamic nature of DNA methylation, as well as the role of transcription factors in this process, is being increasingly recognised (Zhu *et al.*, 2016).

Transcription factors may have direct effects on DNA methylation. Some transcription factors can promote increased DNA methylation by directly recruiting DNMTs. The mouse orphan nuclear receptor 6A1 (Nr6a1), for example, can recruit Dnmt3a, resulting in methylation of the Oct4 promoter during embryonic stem cell differentiation (Sato *et al.*, 2006). Another example is the recruitment of DNMT3A by a MYC / ZBTB17 complex, resulting in the methylation and repression of cyclin-dependent kinase inhibitor 1A (Cdkn1a or p21) (Brenner *et al.*, 2005). An *in vitro* array of 103 transcription factors identified 79 candidate transcription factors interacting with DNMT3A and/or DNMT3B, indicating that sequence-specific transcription factor recruitment of DNMTs may be more widespread than is currently appreciated (Hervouet *et al.*, 2009). However, a significant overlap between transcription factor and DNMT binding sites has not yet been demonstrated in genome-wide ChIP-seq studies (Blattler and Farnham, 2013).

Conversely, other transcription factors have been shown to promote demethylation of DNA, for example FOXA1, CCCTC-binding factor (CTCF) and RE1 silencing transcription factor (REST) (Feldmann *et al.*, 2013; Stadler *et al.*, 2011; Zhang *et al.*, 2016c). Importantly, the binding of these transcription factors and subsequent demethylation is enriched in non-CGI distal regulatory elements (enhancers), suggesting that transcription factors may maintain tissue-specific enhancers in a dynamic low-methylation state through recruitment of demethylating enzymes. A direct interaction between the ETS transcription factor PU.1 and TET2 has been demonstrated during osteoclast differentiation (de la Rica *et al.*, 2013). In addition, deletion of Rest in mouse embryonic stem cells causes increased total methylation at sites of Rest binding but decreased levels of 5hmC, indicating a reduction in TET-mediated turnover of methylation (Feldmann *et al.*, 2013). Consistent with a role in establishing and maintaining enhancers, the ability to initiate DNA demethylation has been proposed to be an important feature of pioneer factors (Domcke *et al.*, 2015; Zhu *et al.*, 2016). However, in contrast to FOXA1, the pioneer factor OCT4 cannot establish

a nucleosome-depleted region in the presence of methylated DNA (You *et al.*, 2011).

The functional consequences of DNA methylation are dependent on the genomic context. Methylation of promoter CGIs is usually associated with stable, long-term silencing, for example genes on the inactive X chromosome in females (Jones, 2012). Methylation of non-CGI promoters (and most likely enhancers) is also generally associated with gene repression, due to inhibition of transcription initiation. This may be a result of direct effects on transcription factor binding (particularly when transcription factors recognise CG-rich motifs) or due to the recruitment of methyl-CpG-binding co-factors, such as the MBD family. Members of the methyl-CpG binding domain (MBD) protein family, including MECP2 (methyl-CpG binding protein 2), MBD1, MBD2 and MBD4, can bind to 5mC, while MBD3 may bind to 5hmC (Rothbart and Strahl, 2014; Yildirim *et al.*, 2011). DNA-bound MBD proteins can recruit histone modifying enzymes and chromatin remodelling complexes (for example, NuRD and the H3K9 methyltransferase SETDB1) to create a repressive chromatin structure (Frauer *et al.*, 2011). Conversely, the maintenance of accessible chromatin at promoter CGIs may be mediated by the action of proteins that specifically recognise unmethylated CpGs through a zinc finger-CXXC domain. Examples include the H3K36 histone demethylase KDM2A, the H3K4 methyltransferase MLL1 and the CXXC zinc finger protein 1 (CXXC1, also known as CFP1) (Blackledge *et al.*, 2010; Cierpicki *et al.*, 2010; Thomson *et al.*, 2010). CXXC1 bound to non-methylated CGIs recruits the SETD1 H3K4 methyltransferase complex (Thomson *et al.*, 2010). As H3K4 methylation inhibits the recruitment of DNMTs, the direct or indirect recruitment of H3K4 methyltransferases by unmethylated CpGs may provide a mechanism by which CGIs are protected from methylation independently of the transcriptional state (reviewed in Cheng, 2014).

As for histone and chromatin modifications, the causal nature of DNA methylation has been widely debated - in other words, does DNA methylation directly cause gene silencing (its primary association) or does it occur after other repressive mechanisms have been initiated. Increasing evidence suggests that DNA methylation is a later event, acting to stabilise the repressed state (Frauer *et al.*, 2011; Jones, 2012). An early finding was that methylation of the mouse X chromosome Hprt gene occurred after genes on the second X chromosome were inactivated in the developing female embryo (Lock *et al.*, 1987). Further evidence comes from the ENCODE project, correlating genome-wide DNA methylation of transcription factor binding sites with transcription factor expression. The strong negative association between binding site

methylation and transcription factor expression indicates that methylation most likely occurs after transcription factors have vacated DNA (in contrast to active eviction of transcription factors by DNA methylation, in which no correlation would be expected) (Thurman *et al.*, 2012). Molecular mechanisms of DNA methylation also point towards methylation as a late event in gene repression. The recruitment of the DNMT3A/3L tetramer in embryonic stem cells, for example, requires a nucleosome and is strongly inhibited by H3K4 methylation; this indicates that mechanisms to compact chromatin and demethylate H3K4 (and subsequently evict protective CXXC factors) are required at active regulatory elements before DNA methylation can occur (Ooi *et al.*, 2007). A similar sequence of events is seen with OCT4 silencing during differentiation of embryonic carcinoma cells, in which a reduction in OCT4 levels triggers increased nucleosome occupancy within the OCT4 enhancer; only once the DNA becomes incorporated into nucleosomes is DNMT3A recruited to methylate the DNA (You *et al.*, 2011). Thus, while methylation can clearly have effects on chromatin structure, histone modifications and transcription factor binding, it seems unlikely that it is the initiating event in gene repression. In fact, the absence of transcription factor binding may even be sufficient to establish regions of nucleosome occupancy followed by DNA methylation.

## 1.2 Regulation of Transcription Factors

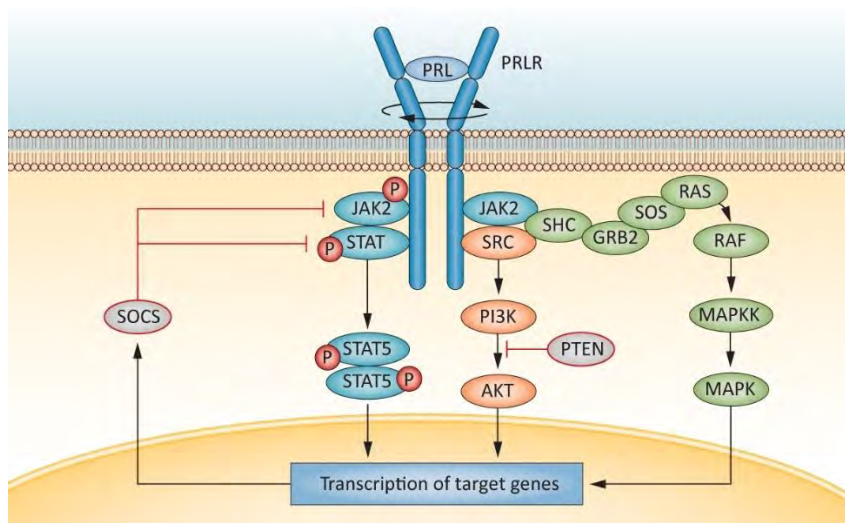
Correct regulation of gene expression by transcription factors is important for normal cellular development and physiology, while misregulation is frequent in diseases such as cancer. Multiple mechanisms therefore exist to regulate transcription factors and ensure their correct spatial and temporal activity. These mechanisms include signalling pathways (culminating in ligand binding or post-translational modifications), protein levels (a balance between expression and degradation), protein-protein interactions, auto-inhibition, subcellular localisation and non-coding RNAs. An additional level of regulation modifies the access of transcription factors to DNA, through the complex interplay of chromatin structure, histone modifications and DNA methylation.

### Signalling Pathways

Transcription factors integrate a variety of signals and transmit this information to the DNA in the form of regulated gene expression. In this way, regulation of gene expression is coupled to internal signals (for example, nutrient levels, DNA damage or viral infection) as well as external signals (for example, hormones, growth factors or temperature). A typical signalling pathway responding to external cues is initiated by the binding of a ligand to a transmembrane receptor (for example, the hormone prolactin binding to the dimerised prolactin receptor (PRLR) long isoform) (Figure 1.9). This initiates a conformational change in the receptor, resulting in the activation of one or more cytoplasmic transducers, which are commonly post-translational modifying enzymes. Ultimately, many signalling transduction pathways culminate in the post-translational modification of a transcription factor, leading to alteration of its activity, binding partners or subcellular localisation (Benayoun and Veitia, 2009). In the case of the canonical prolactin signalling pathway, prolactin binding to PRLR activates Janus kinase 2 (JAK2), which phosphorylates several tyrosine residues within PRLR. This subsequently allows the docking and phosphorylation of signal transducer and activator of transcription A/B (STAT5A/B) proteins, which then homo- or hetero-dimerise and translocate to the nucleus. Binding of STAT5 dimers to DNA results in the activation of target genes, for example milk proteins such as b-casein in the mammary gland. STAT5 also activates the expression of suppressor of cytokine signalling (SOCS) genes, which interact with JAK2 and PRLR to inhibit STAT5 phosphorylation. This establishes a negative feedback loop to limit the transcriptional response to prolactin signalling (Bernard *et al.*, 2015; Hennighausen and Robinson, 2005).



In addition to the major JAK-STAT pathway, other kinase signalling cascades can be activated by prolactin, as shown in Figure 1.9. However, this figure does not capture the full complexity of prolactin signalling. A pathway map of prolactin signalling from a 2013 review, for example, identified 87 protein-protein interactions, 97 enzymatic reactions and 21 protein translocation events that occurred during prolactin signalling (Radhakrishnan *et al.*, 2012). Simultaneous activation and cross-talk between pathways may also occur, generating an elaborate signalling network.



**Figure 1.9: Multiple signalling cascades are initiated on binding of prolactin to the long isoform of the prolactin receptor**

*Figure and caption adapted with permission from Nature Publishing Group (Bernard et al., 2015)*

Binding of prolactin to the dimerised prolactin receptor (PRLR) causes a conformational change (circular arrows), triggering several signalling cascades. The main pathway involves the activation of JAK2, leading to the phosphorylation and dimerisation of STAT5A/B. Dimerised STAT5 then translocates to the nucleus and binds to DNA to activate target genes. STAT5 binding also activates SOCS expression, inhibiting JAK2 and PRLR in a negative feedback loop. Prolactin can also activate the RAS-RAF-MAPK pathway, mediated by the binding of adaptor proteins such as SHC, GRB2 and SOS to PRLR, and the P13K-AKT pathway. AKT, AKT serine/threonine kinase; GRB2, growth factor receptor bound protein 2; JAK2, janus kinase 2; MAPK, mitogen-activated protein kinase; MAPKK, MAPK kinase; PI3K, phosphatidylinositol-4,5-bisphosphate 3-kinase; PRL, prolactin; PTEN, phosphatase and tensin homolog; RAF, rapidly accelerated fibrosarcoma family members (serine/threonine kinases); RAS, rat sarcoma family members (GTPases); SHC, SHC adaptor protein 1; SOCS, suppressor of cytokine signalling; SOS, son of sevenless family members; SRC, SRC proto-oncogene non-receptor tyrosine kinase; STAT5A/B, signal transducer and activator of transcription A/B.

## Ligand Binding

Some transcription factors are directly regulated by ligand binding. The nuclear receptor superfamily contains 48 members, which are characterised by the presence of a structurally conserved ligand binding domain (LBD). A diverse range of small diffusible ligands bind to this domain to regulate receptor activity. These ligands include hormones, retinoic acid, fatty acids, cholesterol, vitamin D and even gases such as nitric oxide and carbon monoxide. Approximately half of all human nuclear receptors, however, remain “orphans”, with the endogenous ligands unknown (Pardee *et al.*, 2011).

The nuclear receptors are unique in their coupling of ligand-binding and DNA-binding abilities. An example of the domain structure of a typical nuclear receptor (GR) is shown in Figure 1.7A, with the DBD and C-terminal LBD connected by a flexible hinge region. The N-terminal domains of nuclear receptors are less well-conserved and are frequently involved in transactivation (Pardee *et al.*, 2011; Weikum *et al.*, 2017). Based on the DBD and LBD sequences, nuclear receptors can be divided into 6 main subfamilies (examples shown in Table 1.2, top). Nuclear receptors can also be divided into 4 classes based on their primary mode of DNA binding (Table 1.2, bottom). Members of class III and IV are mostly orphan receptors and their mechanisms of action are less well understood (Pardee *et al.*, 2011).

Glucocorticoid receptor and oestrogen receptor alpha (ER) are two examples of nuclear receptors regulated by the binding of steroid hormones. In the absence of ligand, GR exists as a monomer in the cytoplasm, bound to molecular protein complexes containing heat shock proteins 70 and 90. Upon binding of endogenous cortisol (or the drug dexamethasone), conformational changes occur in the receptor, leading to dissociation from the chaperone proteins and activation or exposure of functional domains. Ligand-bound GR then translocates to the nucleus, where it interacts with glucocorticoid response elements (GREs). The canonical GR binding sequence (GBS) consists of two pseudo-palindromic hexameric repeats separated by a 3 bp spacer. One GR molecule binds to each repeat in a head-to-head configuration, with interactions between the two DBDs forming a dimerisation interface. GR may also bind DNA as a monomer, either via an inverted-repeat GBS or a single hexameric motif (GBS half-site, regulating transcription with co-operating transcription factors). This flexibility in DNA binding sites is influenced by interactions with co-factors and co-operating transcription factors (Weikum *et al.*, 2017). In the absence of ligand, ER mostly exists as a monomer in the nucleus, although a small amount may be present in

the cytoplasm bound to heat shock proteins. Binding of ligand (for example, oestradiol) promotes the dimerisation and head-to-head DNA binding of ER molecules (Levin and Hammes, 2016). Post-translational modifications such as phosphorylation also modulate the activity of ligand-bound receptors both before and after binding to DNA (Maggi, 2011).

However, the action of nuclear receptors is not always ligand-dependent. ER, for example, can be activated in the absence of ligand by various growth factors, cytokines, other hormones, neurotransmitters and environmental molecules (Stellato *et al.*, 2016). The most common mechanism of ligand-independent activation involves phosphorylation of ER by downstream effectors of these molecules. Important growth factor pathways involved in ER activation are mediated by receptor tyrosine kinases including epidermal growth factor receptor (EGFR) and insulin-like growth factor 1 receptor (IGF1R). These pathways have also been implicated in the ligand-independent activation of ER in endocrine therapy-resistant breast cancer (Musgrove and Sutherland, 2009). Binding of epidermal growth factor to EGFR, for example, triggers dimerisation of the receptor and auto-phosphorylation of key residues. This in turn results in the activation of several kinase signalling cascades, including the RAS-RAF-MAPK and P13K-AKT pathways (The Uniprot Consortium, 2017). Mitogen-activated protein kinase (MAPK) can phosphorylate ER at several key residues, including serine 118 (associated with increased transcriptional competence) and serine 167 (Stellato *et al.*, 2016). However, it has been shown that growth factor-mediated ER activation results in a unique set of ER binding sites in MCF7 breast cancer cells, indicating that the upstream activation signal may alter the transcriptional outcome (Lupien *et al.*, 2010). Similarly, different types of ligand may also result in unique receptor conformations and transcriptional effects (Hall *et al.*, 2002). These differences may occur through modified interactions with DNA or co-factors, consistent with the allosteric model of signal integration previously discussed for GR (Figure 1.7B) (Weikum *et al.*, 2017).

**Table 1.2: The nuclear receptor superfamily**

CLASSIFICATION BASED ON DBD AND LBD SEQUENCE	
Sub-family	Examples
1) Thyroid hormone receptor-like	Liver X receptor (LXR), peroxisome-proliferator-activated receptor (PPAR), RAR-related orphan receptor (ROR), retinoic acid receptor (RAR), Rev-erb, thyroid hormone receptor (TR), vitamin D receptor (VDR)
2) Retinoid X receptor-like	Hepatocyte nuclear factor 4 (HNF4), retinoid X receptor (RXR),
3) Oestrogen receptor-like	Androgen receptor (AR), oestrogen receptor (ER), oestrogen-related receptor (ERR), glucocorticoid receptor (GR), mineralocorticoid receptor (MR), progesterone receptor (PR)
4) Nerve growth factor IB-like	Nerve growth factor IB (NGFIB), neuron-derived orphan receptor 1 (NOR1), nuclear receptor related 1 (NURR1)
5) Steroidogenic factor-like	Liver receptor homolog 1 (LRH1), steroidogenic factor 1 (SF1)
6) Germ cell nuclear factor-like	Germ cell nuclear factor (GCNF)

CLASSIFICATION BASED ON MODE OF DNA BINDING		
Class	Binding mode	Examples
Class I: Steroid homodimers	Ligand binding in cytoplasm, translocation to nucleus, DNA binding to inverted hexameric repeats as homodimers in head-to-head configuration, sometimes bind as monomers	AR, ER, ERR, GR, MR, PR
Class II: Heterodimers	Form heterodimers with RXR in head-to-tail configuration, bind to DNA in absence of ligand to cause gene repression, conformational change on ligand binding results in gene activation	PPAR, RAR, RXR, TR, VDR
Class III: Non-steroidal homodimers	DNA binding to direct hexameric repeats as homodimers	HNF4, RXR
Class IV: Monomers	DNA binding to single hexameric motif (as monomer or dimer)	NGFIB, SF1

References: (Jacobsen and Horwitz, 2012; Pardee *et al.*, 2011; Weikum *et al.*, 2017)

## Post-Translational Modifications

Many signalling pathways convey information about the intra- and extra-cellular environment through post-translational modifications (PTMs) of transcription factors, converting these cues into functional changes in gene expression by altering transcription factor levels, location or activity. Transcription factor PTMs are added to amino acid side chains by specific enzymes and, just like histone modifications, they are reversible. Types of PTMs include phosphorylation, glycosylation, acetylation, methylation, ubiquitination, sumoylation and more recently discovered modifications such as neddylation and ADP-ribosylation (Table 1.3). A recent analysis indicates that most of these PTMs occur in transcription factors at a similar rate to other proteins, with the exception of sumoylation (occurring in published studies at 6.6x compared to non-transcription factors) and ubiquitination (0.47x) (Filtz *et al.*, 2014). PTMs may regulate transcription factors by changing their subcellular localisation, stability, interactions with

co-factors, transcriptional activity or DNA-binding affinity. In addition, many transcription factors can undergo multiple PTMs, which may have combinatorial or sequential effects. The presence of one PTM, for example, may enhance or inhibit the placement of other PTMs (Filtz *et al.*, 2014). Due to the breadth of this topic, this discussion will focus on phosphorylation, followed by an example of a transcription factor regulated by multiple PTMS (p53).

**Table 1.3: Post-translational modifications (PTMs) of transcription factors**

PTM	Definition	Enzymes	Structural effects
Phosphorylation	Addition of a phosphate to the hydroxyl group of a serine, threonine or tyrosine residue	Kinases (+) Phosphatases (-)	Increase in negative charge
Acetylation	Addition of an acetyl group to a lysine residue	Acetylases (+) Deacetylases (-)	Neutralises charge and competes with ubiquitination and sumoylation
Methylation	Addition of a methyl group to a lysine or arginine residue; like histones, lysines can be mono-, di- or tri-methylated, while arginines can be mono- or di-methylated (asymmetric or symmetric)	Methyltransferases (+) Demethylases (-)	Increases effective radius of the positive charge due to addition of the bulky methyl group
Glycosylation	Commonly O-GlycNAcylation, the addition of the monosaccharide $\beta$ -D-N-acetylglucosamine (GlcNAc) to the hydroxyl group of a serine or threonine residue	O-GlcNAc transferase (+) O-GlcNAcase (-)	Does not alter charge, may compete with phosphorylation
Ubiquitination	Addition of a 9 kDa ubiquitin peptide (~76 aa) to a lysine residue (mono-ubiquitination); poly-ubiquitination can occur through various lysines in the ubiquitin molecule, forming ubiquitin chains	E3 ubiquitin ligases (+) Deubiquitinases (-)	Increased bulk, competes with acetylation and sumoylation
Sumoylation	Addition of a 10-11 kDa small ubiquitin-like modifier (SUMO) protein (~100 aa) to a lysine residue (mono-sumoylation); poly-sumoylation can occur via through various lysines in the SUMO molecule, forming SUMO chains	SUMO E3 ligases (+) SUMO-specific proteases (-)	Significantly increases size of protein and bulk of side chains (is the largest single PTM) and competes primarily with ubiquitination
Neddylation	Addition of neural precursor cell expressed developmentally downregulated protein 8 (NEDD8), a ubiquitin-like protein (8 kDa and 81 aa); poly-neddylation may occur	NEDD8 E3 ligases (+) Deneddylases (-)	Increased bulk

Description of the types and structural consequences of transcriptional factor post-translational modifications, and the enzymes that add (+) or remove (-) these modifications (Charlot *et al.*, 2010; Enchev *et al.*, 2015; Filtz *et al.*, 2014; Seeler and Dejean, 2017).

Phosphorylation is the addition of a phosphate to the hydroxyl group of a serine, threonine or tyrosine residue. It is carried out by kinases, which recognise and phosphorylate short amino acid motifs (phosphoacceptor motifs). Additional substrate specificity may be provided by the interaction of kinases with kinase docking domains in transcription factors (Sharrocks *et al.*, 2000). Phosphorylation results in an increase in the negative charge of the modified residue, which can alter protein structure and protein-protein interactions (Holmberg *et al.*, 2002). The functional effects of phosphorylation are therefore dependent on the transcription factor as well as the specific phosphorylation site. Phosphorylation of the ETS1 serine-rich region (SRR) by calcium / calmodulin-dependent kinase II (CaMKII), for example, promotes the auto-inhibited conformation of ETS1, reducing DNA binding (see “Auto-inhibition” sub-section). Conversely, N-terminal phosphorylation of threonine 38 and serine 41 by MAPK1 (which interacts with a docking site in the ETS1 Pointed domain) enhances the interaction with the co-activator CREBBP, increasing ETS1 transcriptional activity (reviewed in Garrett-Sinha, 2013; Hollenhorst *et al.*, 2011). The phosphorylation of the unstructured ETS1 SRR is also an example of how multiple phosphorylation sites can act incrementally to regulate protein activity. Gradual changes in ETS1 structure and DNA binding affinity correlate with the number of phosphorylated SRR sites, thereby providing a precise mechanism for fine-tuning ETS1 transcriptional activity in response to the signal magnitude (Pufall *et al.*, 2005). Additional examples of transcription factors that are regulated by phosphorylation, discussed at other points in the text, include STAT5, STAT1, ER, GR, ETS1 and nuclear factor of activated T-cells (NFAT) proteins.

An example of a transcription factor that is regulated by many types of PTMs at multiple sites is tumour protein 53 (p53), which is essential in the transcriptional response to various cellular stresses. Full-length canonical p53, also known as p53a (393 amino acids), contains two N-terminal transactivation domains (TADs, amino acids 1-42 and 43-92), a central DNA-binding domain (DBD, 102-292), a tetramerisation domain (TD, 326-356) and a C-terminal regulatory domain (CTD, 364-393). Known p53 PTMs, occurring on at least 50 residues, include phosphorylation, acetylation, methylation, ubiquitination, sumoylation, neddylation, glycosylation and ADP-ribosylation (see Kruse and Gu, 2008 for an excellent summary). However, for many of these PTMs, the responsible signalling pathways and functional consequences are unclear (reviewed in Kruse and Gu, 2008; Meek and Anderson, 2009).

The first identified p53 PTM was phosphorylation. A number of N-terminal phosphorylation events within the TADs occur in response to signals such as DNA damage and UV light, stabilising p53 by inhibiting its interaction with the E3 ubiquitin ligase MDM2. Phosphorylation of serine 15, usually by the DNA-damage response kinases ATM and ATR or by AMP-activated kinase (AMPK) in response to glucose depletion, promotes the subsequent phosphorylation of additional N-terminal residues and also recruits the acetyltransferases p300 and CREBBP (Meek and Anderson, 2009). In contrast, phosphorylation of several residues within the DBD occurs in unstressed cells and promotes p53 degradation (Kruse and Gu, 2008).

Acetylation of p53 is also an important regulatory mechanism. As noted above, phosphorylation of p53 promotes the recruitment of the co-factors p300 and CREBBP, which acetylate histones as well as p53. Acetylation of lysine residues in the CTD occurs during cellular stress and increases the stability of p53 (by preventing the ubiquitination of these same sites by MDM2) as well as its transcriptional activity (Meek and Anderson, 2009). DNA damage also promotes the acetylation of p53 at various other lysine residues, including lysines 120, 164 and 320. Acetylation of lysine 120 by KAT5 (TIP60) or KAT8 (MOF) is required for the induction of p53-regulated pro-apoptotic genes but not cell cycle arrest genes (Kruse and Gu, 2008; Meek and Anderson, 2009). Methylation of lysine 372 is a pre-requisite for lysine 120 acetylation (Anderson and Appella, 2010). Interestingly, p300 and CREBBP can also recruit MDM2 to p53, leading to p53 degradation; the balance between transactivation and degradation in this context is largely regulated by the degree of N-terminal phosphorylation, which promotes the p53-acetyltransferase interaction and inhibits the p53-MDM2 interaction (Meek and Anderson, 2009).

The stability of p53 is primarily regulated by the E3 ubiquitin ligase MDM2. In unstressed cells, p53 is polyubiquitinated on multiple lysine residues within the CTD, targeting it for degradation by the proteasome. MDM2 is also a transcriptional target of p53, thereby establishing a negative feedback loop of activation-induced degradation. Mono-ubiquitination by lower levels of MDM2 promotes nuclear export of p53, providing an additional mechanism for controlling p53 activity in unstressed cells (Hammond-Martel *et al.*, 2012). Ubiquitination at other sites by the atypical E4F1 ubiquitin ligase promotes increased p53 transcriptional activation of cell cycle arrest genes (Hock and Vousden, 2014).

The above discussion provides a glimpse of the complex interactions and inter-dependence of PTMs in the regulation of transcription factors. PTMs, and their

combinations, also provide a mechanism by which the same transcription factor can regulate different genes in response to various cellular signals. Acetylation of lysine 120, for example, leads to the specific activation of p53-regulated pro-apoptotic genes, while E4F1-mediated ubiquitination of lysine residues promotes activation of a distinct set of cell cycle arrest genes. This has led to the proposal of a non-histone protein “PTM code” hypothesis, whereby combinations of PTMs specify functional outcomes (Benayoun and Veitia, 2009). However, the intricacies of PTM regulation can also be explained by the allosteric model, in which PTMs provide surfaces that are recognised by specific effector molecules (for example, transcriptional co-factors) in the absence of a general code (Sims and Reinberg, 2008).

### **DNA Binding Sites**

As for ligands, specific DNA sequences can also function as direct allosteric regulators of transcription factors. DNA is not a passive docking site for transcription factors but can directly modulate the conformation of the bound transcription factor in a sequence-dependent manner. This, in turn, may alter the availability or stability of protein surfaces that mediate interactions with co-factors or other transcription factors. In this way, the DNA-binding domain relays sequence-specific signals to other protein domains to drive different transcriptional outcomes (Lefstin and Yamamoto, 1998; Weikum *et al.*, 2017).

An example is the nuclear receptor GR (Figure 7B). There is considerable sequence diversity among the hexameric half-site GR binding sequences (GRSs) that regulate different genes. Only 6 bp in the 15 bp GRS motif, for example, are strongly conserved, corresponding to the 6 bp that make direct contact with GR molecules. The remaining nucleotides are highly variable, with the result that 90% of all GRSs are unique (Watson *et al.*, 2013). Conversely, many GRSs regulating the same target gene are highly conserved across species (Meijsing *et al.*, 2009). This indicates that the precise GRS may convey gene-specific information about the transcriptional response to GR binding at a given site. A similar diversity, combined with gene-specific interspecies conservation, has been identified for the binding sites of nuclear factor kappa B (NF $\kappa$ B) (Leung *et al.*, 2004).

Small changes to the binding site DNA sequence can significantly affect GR conformation and transcriptional activity. In one study, a panel of GRSs, differing by as little as 1 bp, were used to drive expression of a luciferase reporter gene. All the GRSs showed similar baseline activity but were activated from 2-fold to 20-fold by



dexamethasone and these differences were unrelated to variations in GR binding affinity. Structural studies of these GRSs in complex with the GR DBD demonstrated differences in the conformation of the GR lever arm, a loop region within the DBD that does not directly contact DNA but is important for transcriptional activity (Meijsing *et al.*, 2009). Another structural study focused on nucleotide variations in the 3 bp spacer sequence, identifying key structural changes in the lever arm, as well as the DBD dimerisation interface and DNA recognition helix H1. The study also found that allosteric signals could be communicated between GR molecules via the dimerisation interface, meaning that sequence-specific signals from the GBS as a whole are integrated by both GR molecules (Watson *et al.*, 2013). Collectively, these studies indicate that the DNA binding sequence can modify GR conformation at the DNA-binding interface and that these conformational changes can be transmitted to other domains, such as the dimerisation interface. This provides a mechanistic explanation for how DNA-mediated conformational changes may alter interactions with co-operating transcription factors and co-factors.

The allosteric effects of DNA sequence on transcription factors are not limited to GR. Specific ER response elements (EREs), for example, can affect the transcriptional activity of ER by altering the conformation of a co-factor binding pocket in the activation function 2 (AF2) domain of ER. This results in the differential recruitment of co-factors such as steroid receptor co-activator (SRC) proteins 1-3 (Hall *et al.*, 2002). Similarly, the alteration of a single nucleotide in the DNA sequence of an NF $\kappa$ B binding site can change the requirement for the co-factor B cell CLL/lymphoma 3 (BCL-3) without altering NF $\kappa$ B binding affinity (Leung *et al.*, 2004). In mouse pituitary gland lactotropes (expressing prolactin), different binding sequences for the POU domain transcription factor Pit1 result in Pit1-mediated activation of the prolactin (*Prl*) gene and Pit1-mediated repression of the growth hormone (*Gh1*) gene. Pit1, like all POU transcription factors, has a flexible bipartite DBD. A 2 bp increase in the spacer sequence between the binding sites for these two domains in the *Gh1* promoter changes the binding of the POU domains so that they are positioned on the same face of the DNA, compared to perpendicular faces in the *Prl* promoter. Furthermore, the co-repressor NCoR was found to be recruited by Pit1 to the *Gh1* promoter but not to the *Prl* promoter. In this case, the recruitment of different co-factors by the same transcription factor in different conformations results in completely opposite transcriptional outcomes (Scully *et al.*, 2000). These studies provide further evidence that DNA functions as a sequence-specific allosteric regulator of transcription factors belonging to diverse structural families.

## Protein-Protein Interactions

As discussed previously, interactions between transcription factors, co-factors and other proteins are important mechanisms by which regulatory specificity is achieved. Transcription factors can interact with other proteins on DNA (for example, with co-operating transcription factors or co-factors) or outside DNA (for example, with post-translational modifying enzymes). These interactions may regulate various aspects of transcription factor function, including DNA binding, target gene selection, transcriptional activity, subcellular localisation and protein turnover (Li *et al.*, 2000).

Protein-protein interactions of transcription factors on DNA have been discussed in previous sections. Homo-or hetero-dimerisation, for example, may alleviate transcription factor auto-inhibition, while co-operative binding with other transcription factors can direct lineage-specific gene regulation. Transcription factors also interact with a diverse range of co-factors, which are proposed to be the primary effectors of allosteric signals that are integrated by transcription factor conformation (see Figure 1.7B and associated discussion) (Weikum *et al.*, 2017).

Transcription factors may also interact with proteins outside the DNA-bound context. Post-translational modifying enzymes, for instance, may interact with transcription factors both on and off DNA. A previously discussed example is the phosphorylation of ER by MAPK in response to growth factor stimulation, enhancing ER transcriptional activity (reviewed in Stellato *et al.*, 2016). Another group of non-DNA-associated proteins that regulate transcription factor function are inhibitory proteins that sequester transcription factors away from DNA binding sites. The inhibitor of NF $\kappa$ B proteins (I $\kappa$ Bs), for example, bind to NF $\kappa$ B family members in the cytoplasm and mask their nuclear localisation sequences. In response to inflammatory stimuli, such as tumour necrosis factor  $\alpha$  or lipopolysaccharide, the I $\kappa$ B kinase (IKK) complex is activated, phosphorylating the I $\kappa$ Bs. This subsequently leads to their ubiquitylation and degradation by the proteasome, releasing NF $\kappa$ B proteins for translocation to the nucleus and activation of their target genes (reviewed by Perkins, 2007).

## Protein Levels

The quantity of a transcription factor within the nucleus is an important determinant of transcription factor function. Transcription factors interact non-specifically with DNA, scanning the sequence for accessible cognate sites. A greater number of available transcription factor molecules, therefore, increases the likelihood that these cognate sites will be bound, with resulting effects on gene expression. Overall, protein level is a

balance between expression (mRNA synthesis, translation and decay) and protein degradation.

The expression of transcription factors is also regulated by transcription factors. This means that many of the mechanisms that regulate transcription factor activity, for example chromatin structure and DNA methylation, also regulate transcription factor levels. In embryonic stem cells, for example, the pluripotent state is maintained by the core transcription factors OCT4, SOX2 and NANOG. These core transcription factors act together to positively regulate their own expression (through clustered enhancers or “super-enhancers”), establishing a robust, interconnected auto-regulatory loop (Whyte *et al.*, 2013; Young, 2011). The high expression level of these transcription factors in ESCs is therefore a powerful influence on cell fate. This is further demonstrated by the generation of induced human pluripotent stem cells (iPSCs) from differentiated cell types through the overexpression of 3-4 transcription factors, invariably including the core factors OCT4 and SOX2 (Takahashi *et al.*, 2007; Yu *et al.*, 2007). At the same time, genes encoding transcription factors involved in cell differentiation are silenced in ESCs by various mechanisms, including the creation of a repressive chromatin structure by polycomb group complexes (Adam and Fuchs, 2016).

Levels of transcription factors are also influenced by mRNA stability and degradation. There are a number of mechanisms that alter the half-life of mRNA once it is produced, thereby regulating the rate of translation and protein abundance. Some transcription factors, for example Fos proto-oncogene AP-1 transcription factor subunit (FOS) and MYC, contain AU-rich elements (AREs) that modify the rate of mRNA degradation. This is a result of the binding of ARE-binding proteins (ABPs), which can recruit or inhibit mRNA-degrading nucleases. ABPs are, in turn, regulated by post-translational modifications mediated by signalling pathways (Schoenberg and Maquat, 2012). The AREs in the FOS and MYC mRNA molecules interact with the ABP ELAV-like binding protein 1 (ELAVL1, also known as HuR), which stabilises the mRNA; this is an important mechanism of increased expression of FOS and MYC in various cancers (Khabar, 2017). Interactions with microRNAs can also modify mRNA degradation and translation (discussed below) (Khabar, 2017), as can modifications to mRNA such as methylation (N<sup>6</sup>-methyladenosine) (reviewed in Zhao *et al.*, 2017).

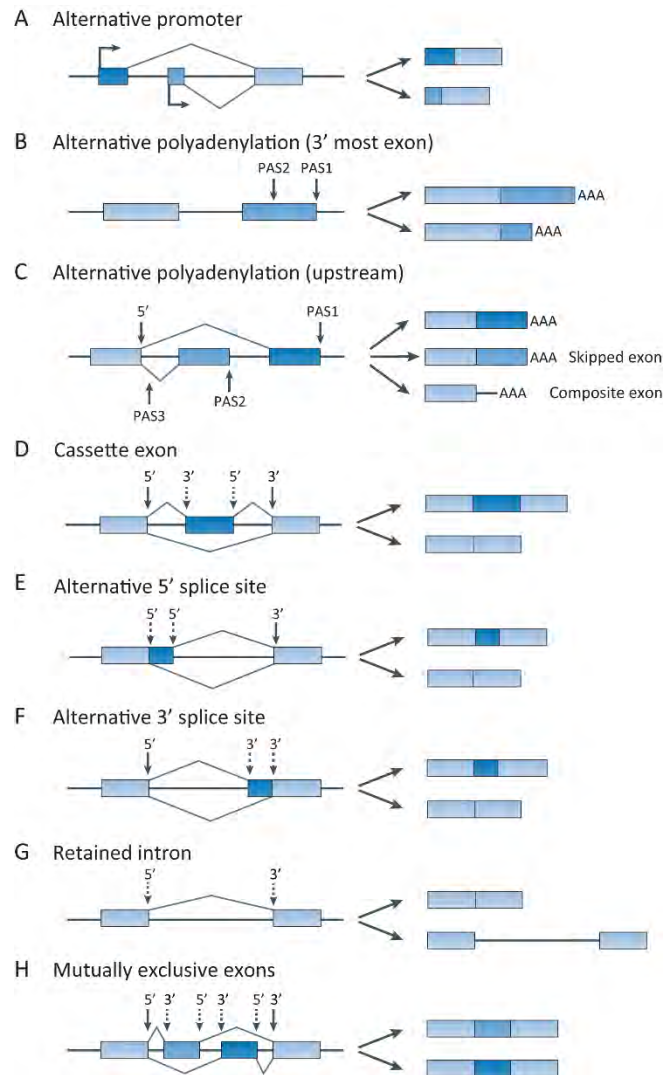
Once protein molecules have been produced from mRNA, the overall abundance is fine-tuned by the rate of protein degradation. For a number of transcription factors, post-translational modifications and co-factor interactions that stimulate transcription factor activity also promote transcription factor degradation, limiting the extent of the

transcriptional response. Examples of transcription factors regulated in this way include NF $\kappa$ B, JUN, MYC and the nuclear receptors oestrogen receptor (ER), androgen receptor (AR), progesterone receptor (PR) and thyroid hormone receptor (TR) (reviewed in Zhou and Slingerland, 2014). A number of ER co-activators are ubiquitin ligases, for example ubiquitin protein ligase E3A (UBE3A), E3 protein-ubiquitin ligase MDM2 (MDM2), and breast cancer type 1 susceptibility protein (BRCA1). The binding of UBE3A is facilitated by the SRC-mediated phosphorylation of ER. Mono- or poly-ubiquitination of ER by UBE3A and other ubiquitin ligases increases ER activity, while simultaneously recruiting the proteasomal machinery to degrade ligand-activated ER (reviewed in Zhou and Slingerland, 2014).

### **Transcript Variants**

Almost all human genes, including transcription factors, undergo alternative transcription and/or alternative splicing, increasing the diversity of protein structure and function. Alternative transcription events arise from the use of alternative transcription start or polyadenylation sites, producing multiple pre-mRNAs. Conversely, in alternative splicing, the pre-mRNAs produced are identical, with co- and post-transcriptional splicing events producing different mature mRNAs (Pal *et al.*, 2012). About 30% of human genes have multiple first exons due to alternative transcription start sites (TSSs), 70% have multiple polyadenylation sites and 90% undergo alternative splicing, producing an average of 6.3 transcripts (or 3.9 protein-coding transcripts) per gene (Figure 1.10) (Djebali *et al.*, 2012; Manning and Cooper, 2017).

Alternative TSSs produce transcripts with different 5' sequences. The use of alternative TSSs is highly tissue-specific and may be regulated by alternative promoters and/or alternative enhancers (Figure 1.10A). Mechanisms controlling the choice of TSS include chromatin structure and the expression of cell-specific transcription factors (de Klerk and 't Hoen, 2015). The use of alternative TSSs may affect the 5' protein-coding sequence, producing proteins with distinct N-termini. In some cases, only the 5' untranslated region (UTR) may be affected, leaving the protein-coding sequence unchanged. However, these changes can still have significant effects on subsequent regulatory events (de Klerk and 't Hoen, 2015). GR, for example, encoded by the nuclear receptor subfamily 3 group C member 1 (*NR3C1*) gene, has at least 9 alternative first exons, which are regulated by unique promoters but do not contribute to protein coding. Overexpression of GR transcripts containing different first exons revealed differences in mRNA secondary structure, mRNA stability, translational efficiency and translational start site (AUG codon) selection (Turner *et al.*, 2014).



**Figure 1.10: Transcript variants are produced by alternative promoters, alternative polyadenylation sites and alternative splicing**

Schematic representation of alternative transcription (A-B) and splicing (D-H) events. On the left, exons are shown as boxes, introns as straight lines and splicing events as angled lines. The structures of the mature mRNA transcripts are shown on the right. Constitutive sequences, which always form part of the mature mRNA, are shown in light blue. Alternative sequences, which are variably included as a result of alternative transcription or splicing events, are shown in mid-blue or dark blue. Constitutive splice sites (marked 5' or 3') are indicated by solid vertical arrows and alternative splice sites by dashed arrows.

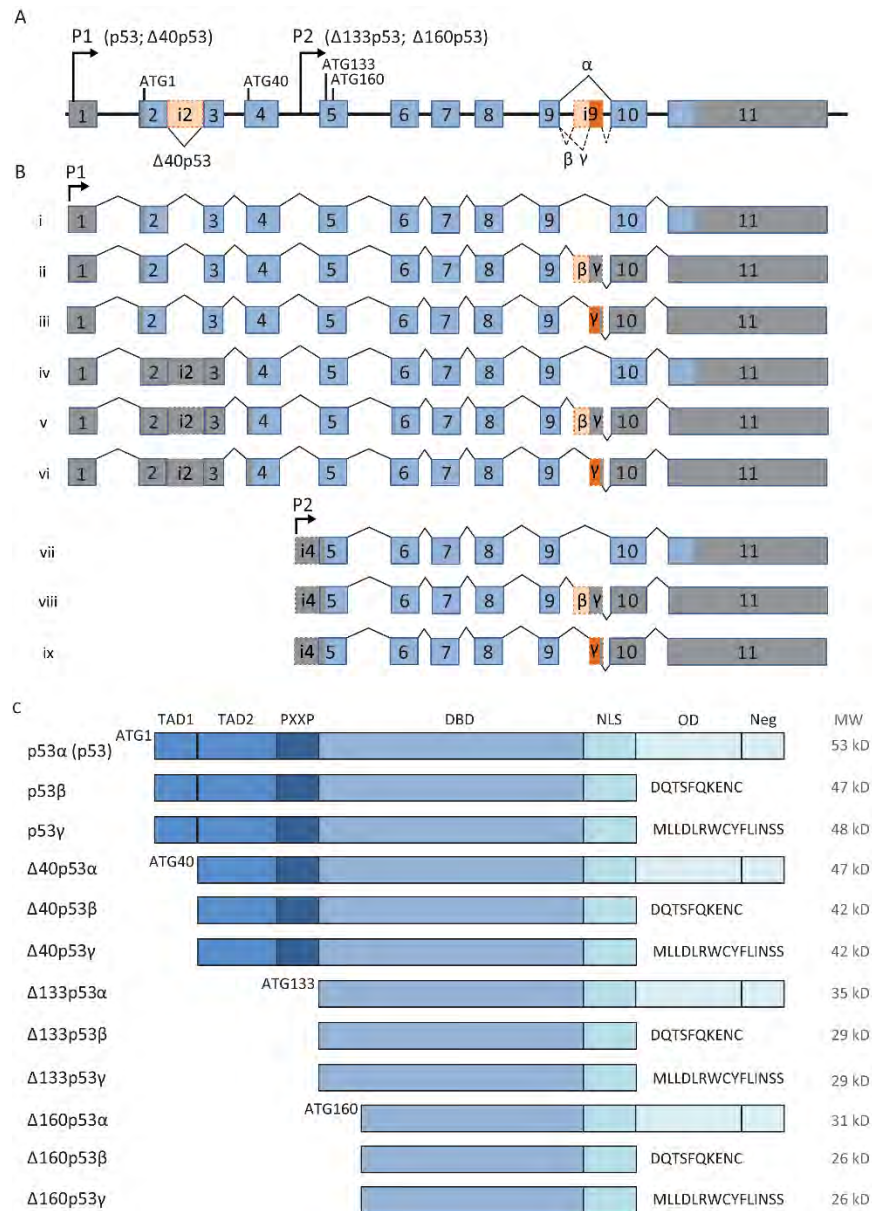
(A) Alternative promoter usage results in different transcription start sites and unique first exons. (B) Alternative polyadenylation and cleavage sites (PAS) within the 3'-most exon results in transcripts with differences in 3' UTR length. The transcript with the longer 3' UTR results from the use of PAS1, while the transcript with the shorter 3' UTR results from the use of PAS2. (C) Alternative PAS in upstream exons or introns results in transcripts with differences in the 3' UTR and in some cases the protein-coding sequence. The 'canonical' transcript is formed by the use of PAS1 in the 3'-most exon. The use of PAS2 in an alternative exon (excluded from the canonical transcript) results in a skipped terminal exon.

The use of PAS3 in an upstream intron (for example, due to inhibition of the indicated 5' splice site) results in inclusion of part of the intron in the final transcript (composite exon). (D) The alternative (cassette) exon is associated with weak or inhibited 3' and 5' splice sites. These splice sites are inefficiently recognised by the spliceosome, resulting in variable usage (indicated by the dashed arrows) and therefore variable exon inclusion. This is the most common type of alternative splicing event in vertebrates. (E) The use of alternative 5' splice sites results in changes in 5' exon length. (F) The use of alternative 3' splice sites results in changes in 3' exon length. (G) Inefficient recognition of the 5' and 3' splice sites defining the intronic region results in intron retention in the final transcript. This is the least common splicing events in vertebrates. (H) Pairing of alternative splice sites results in a pattern of mutually exclusive exon inclusion, as occurs for FOXP1 (see text). 3', 3' splice site; 5', 5' splice site; AAA, poly(A) tail; PAS, polyadenylation and cleavage site. References: (Chen and Weiss, 2015; Dvinge *et al.*, 2016; Naftelberg *et al.*, 2015).

Effects of the 5'UTR on translation have also been demonstrated for other transcription factors, such as the ETS transcription factor ELK1 (Araud *et al.*, 2007; Rahim *et al.*, 2012). General features of the 5' UTR that can influence translation include length, mRNA structure, upstream AUGs, upstream open reading frames and internal ribosome entry site (IRES) elements (Rahim *et al.*, 2012). Therefore, transcript variants arising from alternative transcript start sites may have unique mRNA characteristics that subsequently modulate protein production.

Alternative TSSs may also alter the N-terminal protein-coding sequence, with potential effects on protein function. An example of a transcription factor with N-terminal variants is tumour protein 53 (p53), encoded by the 11-exon *TP53* gene. The human *TP53* gene gives rise to 9 mRNA transcripts encoding 12 protein isoforms with distinct N-termini, produced by combinations of alternative TSSs, alternative splicing and alternative initiation of translation (Figure 1.11). Transcription can be initiated from one of two main promoters (P1, upstream or P2, within intron 4). Transcriptional initiation from P1 produces canonical p53 and D40p53 (which lacks the first 39 amino acids due to the alternative splicing of intron 2 and/or the use of an alternative translation initiation site). Transcriptional initiation from P2 produces D133p53 and/or D160p53 from a single mRNA, with D160p53 arising from the use of a second translation initiation site. Each of these N-terminal variants can also undergo at least three different splicing events (a, b, g) involving intron 9 at the C-terminus, producing a total of 12 protein isoforms with unique domain structures (Surget *et al.*, 2013). The transcript variants arising from the alternative P2 promoter (D133p53 and D160p53) lack the two N-terminal transactivation domains and part of the DNA-binding domain. This has the potential to alter co-factor interactions that are essential for p53 transcriptional activity.

Indeed, most studies indicate that D133p53 inhibits full-length p53 activity, although this may be dependent on the cellular context and the balance of isoforms expressed (reviewed in Chen and Weiss, 2015; Jorruiz and Bourdon, 2016). However, despite the wealth of knowledge about p53, the functional roles of these isoforms, and their cellular interactions, are still being investigated.



**Figure 1.11: The human *TP53* gene encodes 9 mRNA transcripts, producing 12 protein isoforms with unique domain structures**  
**Figures 11A and 11C adapted with permission from Dove Medical Press (Surget et al., 2013) and Figure 11B based on a figure from (Jorruiz and Bourdon, 2016)**

(A) *TP53* gene structure. The *TP53* gene contains 11 exons. Grey exons are always non-

coding and orange introns can be alternatively spliced. Transcription can be initiated from promoter 1 (P1), with translation from ATG 1 (canonical p53) or ATG 40 (D40p53). Transcription can also be initiated from an intronic promoter 2 (P2), with translation from ATG 133 (D133p53) or ATG 160 (D160p53). Alternative splicing of intron 9 produces the C-terminal a, b and g variants. The D40p53 protein isoform can also be produced from a transcript that retains intron 2. (B) p53 mRNAs (labelled i-ix). Grey regions are non-coding and blue/orange regions are protein-coding. Transcripts i-iii can give rise to canonical p53 and/or D40p53 protein isoforms (a, b and g variants). Transcripts iv-vi give rise to D40p53 isoform only (a, b and g variants). Transcripts vii-ix give rise to D133p53 and/or D160p53 isoforms (a, b and g variants). (C) p53 protein isoforms. Protein domains are labelled. ATG, start codon; DBD, DNA-binding domain; i, intron; NLS, nuclear localisation signal; Neg, negative-regulation domain; OD, oligomerisation domain; PXXP, proline-rich domain; TAD, transactivation domain (1 and 2).

Another alternative transcription event is the use of alternative polyadenylation (APA) sites, which produces mRNA transcripts with different 3' ends (Figure 1.10B-C). In a process linked to transcription termination, the nascent RNA is cleaved and a poly(A) tail (a stretch of adenine bases) is added to the 3' end by a poly(A) polymerase. Cleavage and polyadenylation (hereafter referred to as polyadenylation) sites are defined by upstream and downstream DNA sequences, which act as binding sites for the polyadenylation machinery. Upstream sequences are commonly variants of the hexamers AAUAAA or AUUAAA (also known as the polyadenylation signal) or other U-rich or UGUA sequences. Downstream elements include U-rich and GU-rich (often GUGU) sequences, with a CA sequence frequently found immediately 5' of the cleavage site. The 3' end of a gene can encode a number of these sequence elements, thereby generating alternative polyadenylation sites (reviewed in Tian and Manley, 2017). The use of APA sites can produce transcripts that differ in the length of the 3' UTR only (if APAs are located in the most 3' exon) or in the protein-coding sequence and 3' UTR (if APAs are located in upstream exons or introns). Differences in the 3' UTR may affect transcript localisation, stability, translation efficiency and even localisation of the (identical) translated proteins. In general, a longer 3' UTR is associated with a reduced transcript level, which may be related to factors such as the presence of microRNA binding sites or destabilising AU-rich elements in the longer UTR (de Klerk and 't Hoen, 2015; Tian and Manley, 2017). APA site usage appears to be tissue-specific, with many cell types showing preference for certain 3' UTR lengths; neuronal cells, for example, tend to use distal APAs. APA usage is also dynamic during cell growth and development. During differentiation of myoblasts into myotubes, for example, there is a global increase in the 3' UTR length; conversely, when iPSCs are



generated from differentiated cells there is a decrease in 3' UTR length. Regulation of APA usage may involve the expression levels of polyadenylation machinery components (which, for example, may promote the use of proximal or distal APA sites), interactions with the splicing and general transcriptional machinery, the transcriptional elongation rate and chromatin structure (reviewed in de Klerk and 't Hoen, 2015; Tian and Manley, 2013, 2017).

An example of a transcription factor regulated by APA is the myogenic regulator paired box 3 (PAX3). Protein levels of murine Pax3 in quiescent adult muscle stem cells from different anatomical locations have been shown to be regulated by the production of transcripts with unique 3' UTRs. The short 3' UTR forms of Pax3 lack a microRNA 206 (miR-206) binding site and are higher in adult stem cell populations with high Pax3 protein expression (for example, muscles of the diaphragm, ventral body wall and selected limb muscles). Conversely, the long 3' UTR form of Pax3 is degraded by miR-206 targeting and is higher in populations with absent Pax3 protein expression (for example, most hindlimb muscles). Alternative polyadenylation sites therefore provide a molecular mechanism for heterogeneity in the adult muscle stem cell population, with potential functional consequences on myogenic development (Boutet *et al.*, 2012).

The final and most prevalent method for generating transcript variants is alternative splicing, which occurs in more than 90% of human genes. Splicing is the removal of introns from the pre-mRNA. It is performed by the spliceosome, a RNA-protein complex containing the small ribonucleoproteins (snRNPs) U1, U2, U4, U5 and U6, as well as many other protein factors. The beginning and end of each intron are marked by splice sites, which are consensus sequences recognised by the spliceosome snRNPs. The spliceosome assembles at these splice sites, which may be “strong” (closely resembling the consensus sequence and efficiently recognised) or “weak” (less efficiently recognised). Alternative splicing is the differential inclusion of exons in the processed mRNA transcript, arising from the differential use of splice sites. The final splicing pattern is strongly influenced by the relative positions of competing strong and weak splice sites along the RNA molecule. Strong splice sites tend to be constitutively used, while weak sites are not efficiently recognised and are variably used, leading to different patterns of alternative splicing (Figure 1.10D-H). The use of splice sites is regulated by RNA-binding proteins (for example, the serine-arginine-rich and the heterogeneous ribonucleoprotein families) that recognise splicing enhancer and silencer sequences in the pre-mRNA molecule to activate or inhibit splicing (Kornblihtt *et al.*, 2013).

Transcription and splicing are intimately linked, as most splicing events occur co-transcriptionally. Transcription may regulate splicing through direct recruitment of splicing factors by the transcriptional machinery (recruitment coupling) and/or alteration of the rate at which the pre-mRNA molecule including its splice sites and regulatory sequences are produced (kinetic coupling). A faster transcriptional elongation rate, for example, increases the rate of exon skipping (decreased splicing), while a slower elongation rate promotes the inclusion of alternative exons (increased splicing). However, the effect of the elongation rate on splicing may be determined by the balance of regulatory factors involved for a particular alternative exon, as slow elongation in some contexts may allow time for the recruitment of inhibitory splicing regulators, favouring exon skipping. Splicing is also affected by RNA secondary structure (which may, for example, hide or expose splicing regulatory sequences), chromatin accessibility, nucleosome positioning, histone modifications and gene body DNA methylation (Kornblihtt *et al.*, 2013; Naftelberg *et al.*, 2015). The binding of some transcription factors, for example the ETS factor Pu.1, has also been shown to modulate splicing, possibly through direct interactions with splicing factors or effects on the elongation rate (Guillouf *et al.*, 2006). The regulation of splicing, therefore, involves a complex interplay between splicing factor levels and activities, RNA regulatory elements, transcription factors, transcriptional elongation rate, chromatin features and DNA modifications, making prediction of splicing patterns difficult (de Klerk and 't Hoen, 2015). It has been proposed, however, that the regulation of splicing may be as important as the regulation of transcription in the specification of cell lineage and response to external signals (Naftelberg *et al.*, 2015).

Alternative splicing may produce mRNA transcripts with different stabilities (as described above) and/or protein-coding sequences. Protein isoforms arising from splice variants can have different activity or functions due to, for example, changes in protein domains, post-translational modification sites, subcellular localisation or interactions with co-factors. There are many examples of transcription factors with modified activity arising from splice variants (see Kelemen *et al.*, 2013 for an excellent review). One example is the ESC-specific isoform of forkhead box P1 (FOXP1-ES), generated by the inclusion of an alternative exon 18, termed exon 18b (an example of mutually exclusive exons, figure 1.10H). Inclusion of exon 18b instead of exon 18 alters the DNA-binding domain, substituting two key amino acid residues that form direct contacts with DNA. This results in the recognition of a different DNA-binding sequence by FOXP1-ES and the regulation of a distinct set of target genes, which includes activation of the core ESC regulators OCT4 and NANOG (Gabut *et al.*, 2011). Thus, FOXP1-ES is uniquely

expressed in ESCs and is pivotal in the transcriptional regulation of the stem cell state. The splicing factors that regulate exon 18/18b inclusion, which have not yet been identified, are also likely to be essential in the maintenance of pluripotency.

ETS1 is another example of an alternatively spliced transcription factor. In this case, alternative splicing of canonical ETS1 (p51 or p54) produces a variant lacking exon 7 (p42) and a variant lacking exons 3-6 (p27). In the p42 isoform, one of the two auto-inhibitory domains is removed, resulting in unique DNA-binding and transcriptional activity compared to p51 (reviewed in Garrett-Sinha, 2013; Laitem *et al.*, 2009). In the p27 isoform, a domain present in a subset of ETS factors known as the Pointed domain is removed, along with the transactivation domain. The p27 isoform, which binds to DNA but is transcriptionally inactive, inhibits the transcriptional activity of the full-length p51 isoform by competing for DNA-binding sites. It also promotes the translocation of p51 from the nucleus into the cytoplasm (Laitem *et al.*, 2009). This is an example of dominant-negative regulation of a transcription factor mediated by the ratio of the isoforms expressed.

Alternative promoters, polyadenylation sites and splicing greatly increase the transcriptomic and proteomic outputs of complex genomes. As discussed above, with a focus on transcription factors, these processes also provide additional levels of regulation for both RNA and protein molecules. New technologies such as RNA-sequencing are providing an increased understanding of the extent and complexity of these events within the human genome.

### **Auto-inhibition**

A commonly used regulatory mechanism is the inhibition of protein activity by an internal negative control region. These regions may inhibit transcription factor functions such as DNA binding, transcriptional activity or interactions with co-factors. A number of regulatory mechanisms may enhance or relieve transcription factor auto-inhibition, including post-translational modifications, protein-protein interactions or alternative splicing (Garvie *et al.*, 2002).

Multiple members of the ETS transcription factor family are regulated by auto-inhibition (Hollenhorst *et al.*, 2011). ETS1, for example, contains an auto-inhibitory helical bundle (HI-1, HI-2, H4, H5), which interacts with helix H1 of the DNA-binding (ETS) domain. An adjacent serine rich region (SRR) also transiently interacts with the ETS domain and inhibitory helices. Structural studies indicate that these regions inhibit high-affinity ETS1 DNA-binding by promoting a more rigid conformation of the ETS domain, rather

than by physically blocking the interaction with DNA (Garvie *et al.*, 2002; Lee *et al.*, 2005a). The SRR can be phosphorylated at multiple sites, increasing the interaction with the ETS domain. This promotes the more rigid state and enhances auto-inhibition in a graded manner that is correlated with the number of phosphorylated residues (Cowley and Graves, 2000; Hollenhorst *et al.*, 2011; Pufall *et al.*, 2005). Conversely, auto-inhibition may be relieved by the interaction of ETS1 with other transcription factors on DNA. Interactions between ETS1 and DNA-bound RUNX1, for example, result in the displacement of the ETS1 HI-1 and HI-2 helices, dramatically reducing auto-inhibition even in the presence of SRR phosphorylation (Shrivastava *et al.*, 2014). Thus, auto-inhibition provides a mechanism by which transcription factor activity can be regulated in a signal-responsive and cell type-specific manner.

### **Subcellular Localisation**

Transcription factors, in order to regulate transcription, need to be localised to the nucleus. Therefore, mechanisms that modify subcellular localisation are important regulators of transcription factor activity.

The nucleus is surrounded by a lipid bilayer (the nuclear envelope), which contains macromolecular nuclear pore complexes (NPCs) that facilitate the passage of proteins in and out of the nucleus. Proteins less than ~40 kilodaltons (kDa) can generally diffuse through the NPCs, while larger proteins are shuttled through by nuclear transport receptors belonging to the b-karyopherin family (importins and exportins). Karyopherins or adaptor proteins recognise nuclear localisation signals (NLSs) and nuclear export signals (NESs) on the surfaces of proteins, which are then transported through the NPC in a directional process powered by the GTPase RAN (Nardozzi *et al.*, 2010).

Most transcription factors contain at least one NLS. However, the subcellular localisation of transcription factors may be regulated by various factors, including nuclear export sequences, post-translational modifications arising from signalling cascades, splice variants (which may gain or lose NLSs or NESs) or interactions with inhibitory proteins (for example, NF $\kappa$ B proteins with I $\kappa$ Bs, discussed earlier). Some examples of transcription factors that are regulated by changes in subcellular location include GR (Weikum *et al.*, 2017), NF $\kappa$ B (Perkins, 2007), STAT proteins (Meyer and Vinkemeier, 2004), p53 (O'Brate and Giannakakou, 2003), nuclear factor of activated T-cells (NFAT) proteins (Nardozzi *et al.*, 2010), and the ETS transcription factors ELK1 and ELK3 (Charlot *et al.*, 2010).

Post-translational modifications are an important modulator of subcellular localisation. Phosphorylation, for example, may increase nuclear import by increasing the affinity of a protein for importins or by unmasking a nuclear localisation sequence. The transcription factor STAT1 is phosphorylated by Janus kinases (JAKs) in response to extracellular signals such as cytokine binding. Tyrosine phosphorylation triggers STAT1 dimerisation, exposing a dimer-specific NLS (dsNLS) in the DNA-binding domain. The dsNLS interacts with importin  $\alpha$ 5, transporting the phosphorylated dimer into the nucleus where it binds to DNA and regulates cytokine-induced gene expression. Importantly, the dsNLS is only active in the context of the phosphorylated STAT1 dimer (reviewed in Meyer and Vinkemeier, 2004; Nardozzi *et al.*, 2010). However, phosphorylation may also decrease nuclear import by preventing recognition of an NLS by importins. An example is the nuclear factor of activated T-cells (NFAT) family of transcription factors. NFAT proteins are normally located in the cytoplasm due to the combination of a strong NES and phosphorylation of a serine-rich region, which overlaps the NLS. Calcium signalling, resulting in increased intracellular calcium levels, leads to activation of the serine phosphatase calcineurin, which progressively dephosphorylates NFAT at multiple sites. This exposes the NLS and enables transport of NFAT into the nucleus (reviewed in Filtz *et al.*, 2014; Nardozzi *et al.*, 2010).

Importantly, however, not all transcription factors are inactive in the cytoplasm. The ETS transcription factor E74 like factor 3 (ELF3), for example, has a non-transcriptional cytoplasmic function that is mediated through its serine- and aspartic acid-rich (SAR) domain. In breast cancer, ELF3 is overexpressed and is primarily located in the cytoplasm. The cytoplasmic overexpression of ELF3 in mammary epithelial cells can trigger transformation, as demonstrated by enhanced anchorage-independent growth, and this transforming potential is inhibited by deletion of the SAR or by fusion to the SV40 NLS. One NLS and two NES have been identified in ELF3 that mediate its nucleocytoplasmic shuttling. However, the mechanism by which the SAR domain of ELF3 performs these unique non-transcriptional functions is currently unknown (Prescott *et al.*, 2004; Prescott *et al.*, 2011).

## **Non-Coding RNA**

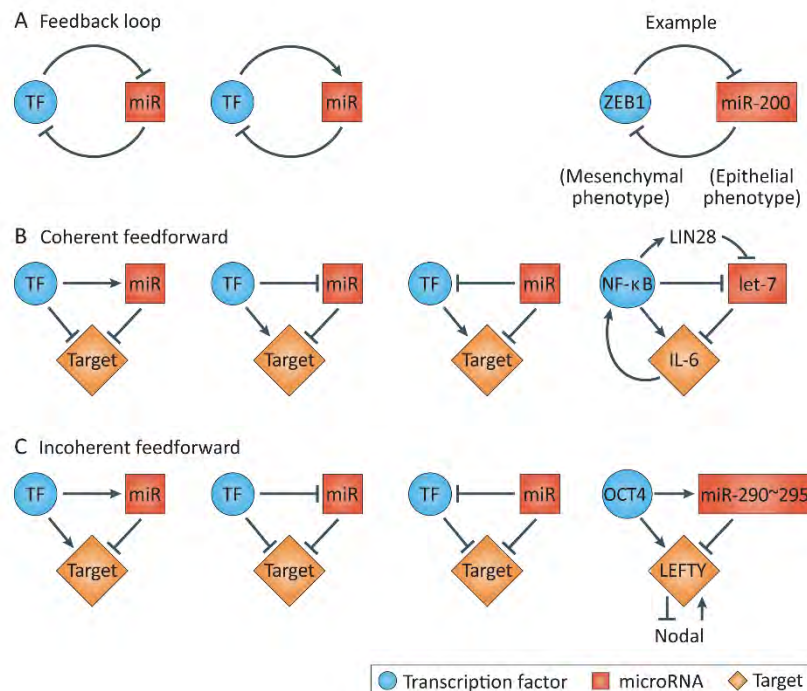
Exons of protein-coding genes account for less than 3% of the human genome (Dunham *et al.*, 2012). However, transcription is not limited to these protein-coding regions; many regions of the genome are in fact transcribed (or differentially transcribed), producing a wealth of non-protein-coding RNA molecules with diverse regulatory functions. There are a number of types of non-coding RNAs, including

microRNAs (miRNAs), small nuclear RNAs, small nucleolar RNAs, PIWI-interacting RNAs, transcription initiation RNAs, splice site RNAs and long non-coding RNAs (lncRNAs, including enhancer, promoter-associated, gene-body-associated and long intervening RNAs) (Bonasio and Shiekhattar, 2014; Morris and Mattick, 2014). This discussion will focus on two types of ncRNAs (miRNAs and lncRNAs) that have known or emerging roles in the regulation of transcription and transcription factors.

MicroRNAs are small RNA molecules (typically 19-25 nucleotides in length) that regulate the post-transcriptional expression of target genes. Primary miRNAs are typically transcribed from intronic or intergenic regions. These molecules are subsequently cleaved and transported to the cytoplasm, where they are processed into double-stranded miRNAs by the endonuclease Dicer. After processing, one strand of the miRNA binds to argonaute proteins within the RNA-induced silencing complex (RISC). Partial base pairing of the miRNA (particularly the 5' seed region) with a target mRNA molecule (typically the 3' UTR) results in decreased mRNA translation and increased degradation (Bracken *et al.*, 2016; Shenoy and Blelloch, 2014). A single miRNA can target hundreds of mRNA molecules and, similarly, a single mRNA molecule can be regulated by many individual miRNAs. In many cases, however, the effects on mRNA levels are modest. The broad regulation of critical cellular processes by miRNAs therefore arises from the small yet simultaneous effects on many genes, combined with the frequent targeting of key transcription factors (Bracken *et al.*, 2016).

Many miRNAs target transcription factors to propagate their effects on gene expression. In fact, a recent study demonstrated that global miRNA depletion mainly influences gene expression result through indirect effects on transcription (via regulation of transcription factors), rather than through direct post-transcriptional effects (Gosline *et al.*, 2016). Transcription factors may also regulate the expression of miRNAs, establishing regulatory loops that include reciprocal feedback loops (transcription factor and miRNA regulate each other's expression), coherent feed-forward loops (transcription factor and miRNA regulate a common target in the same direction) and incoherent feed-forward loops (transcription factor and miRNA regulate a common target in the opposite direction (Figure 1.12) (Bracken *et al.*, 2016). The transcription factors zinc finger E-box binding homeobox 1 and 2 (ZEB1 and 2), for example, are regulated by the microRNA miR-200. ZEB1 and ZEB2 promote the mesenchymal phenotype by repressing the expression of epithelial genes such as E-cadherin. In normal epithelial cells, miR-200 binds to multiple sites in the 3' UTR of ZEB1, limiting its expression. However, ZEB1 expression may be increased by certain

signals (for example, transforming growth factor- $\beta$  (TGF- $\beta$ ) stimulation), leading to ZEB1-mediated inhibition of miR-200 expression. This further increases ZEB1 levels and promotes epithelial to mesenchymal transition (EMT) in an example of a reciprocal feedback loop (Figure 1.12A) (Bracken *et al.*, 2016). miR-200 also targets other genes involved in EMT, such as SMAD family member 2 (*Smad2*, transcription factor also activated by TGF- $\beta$  that can act co-operatively with Zeb), snail family transcriptional repressor 1 (*Snai1*) and the tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation proteins beta and gamma (*Ywhab*, *Ywhag*, co-factors for Snai1) (Perdigao-Henriques *et al.*, 2016). This demonstrates the ability of miRNAs to target multiple components of a gene regulatory network.



**Figure 1.12: Common microRNA-transcription factor regulatory loops**

**Figure reproduced (and caption modified) with permission from Nature Publishing Group (Bracken *et al.*, 2016)**

Three major classes of microRNA-associated signalling feedback loops are represented. Prominent known examples of each class are shown on the left. (A) Direct reciprocal feedback loop in which a transcription factor and microRNA regulate each other's expression. An example is ZEB1 and miR-200, discussed in the text. (B) Coherent feed-forward loop, in which a transcription factor and miRNA regulate a common target gene in the same direction (either activating or repressing). (C) Incoherent feed-forward loop, in which the transcription factor and miRNA regulate a common target in the opposite direction, resulting in a buffering effect. IL-6, interleukin 6; LEFTY, left-right determination

factor; let-7, let-7 miRNA; LIN28, lin-28 homolog A protein; miR, microRNA; Nodal, nodal growth differentiation factor; NF- $\kappa$ B, nuclear factor kappa B; OCT4, POU class 5 homeobox 1 transcription factor; TF, transcription factor; ZEB1, zinc finger E-box binding homeobox 1.

Another important class of non-coding RNAs is long non-coding RNAs (lncRNAs). These are RNA transcripts longer than 200 nucleotides that do not encode proteins. Similar to mRNAs, they are transcribed by RNA Pol II and, in some cases, may be 5'-capped, spliced and polyadenylated. However, in contrast to mRNAs, lncRNAs tend to be shorter, have fewer but longer exons and lack a protein-coding open reading frame (ORF) (Quinn and Chang, 2016). Around 60-70% of lncRNAs are upstream antisense transcripts that originate near the TSS of genes (promoter-associated RNAs). The remaining lncRNAs arise from enhancer regions (eRNAs, ~20%), gene bodies (gene-body-associated RNAs, ~5%), or more distal and currently unannotated regions (long intervening RNAs or lincRNAs, ~5%) (Bonasio and Shiekhattar, 2014). Nuclear lncRNAs primarily regulate transcription by guiding chromatin modifiers to specific genomic loci, either in *cis* (close to where they are transcribed) or in *trans* (at independent loci). In both cases, however, the mechanisms for targeting lncRNAs to genomic sites are not well understood (Fatica and Bozzoni, 2014). Some lncRNAs may interact directly with transcription factors, recruiting additional activating or repressive co-factors or functioning as decoys that sequester transcription factors away from their binding sites. The pluripotency-associated lncRNAs ES1 and ES2, for example, interact with the transcription factor SOX2 and the Polycomb protein SUZ12 to co-ordinate the repression of differentiation-associated genes (Bonasio and Shiekhattar, 2014; Fatica and Bozzoni, 2014). Another important function of lncRNAs, particularly eRNAs, is to guide looping between enhancers and promoters. An enhancer for the Kallikrein-related peptidase 3 (KLK3) gene, for example, produces an eRNA that interacts with both androgen receptor and Mediator, facilitating the interaction between the enhancer the KLK2/3 promoters to increase AR-mediated gene expression (Hsieh *et al.*, 2014). A similar requirement for eRNAs in ER-mediated transcription has also been demonstrated (Li *et al.*, 2013). Some lncRNAs also function as competitive endogenous RNAs (ceRNAs), which bind to miRNAs to inhibit their function, thereby indirectly regulating transcription factor function (Fatica and Bozzoni, 2014). In addition, lncRNAs may affect protein translation; the lncRNA linc-ROR (regulator of reprogramming), for example, inhibits the translation of p53 in the cytoplasm (reviewed in Grossi *et al.*, 2016). The role of lncRNAs in the regulation of gene expression is clearly complex and there is still much to learn about this emerging class of diverse regulatory molecules.



## Chromatin Structure

The structure of chromatin is intrinsically repressive to transcription factor binding. This is an important regulatory mechanism, as the average 8 bp recognition sequence of a transcription factor will occur about 45,000 times in a human-sized genome of random sequence. The number of potential binding sites is in fact likely to be much greater than this estimate due to the ability of transcription factors to bind many sequences with varying affinities (van Bakel, 2011). Chromatin therefore provides a mechanism for controlling the access of transcription factors and the core transcriptional machinery to DNA, ensuring that transcription does not commence from unwanted sites.

The regulated access of transcription factors to chromatin is determined by the interplay of several factors, including DNA sequence, histone modifications and variants, transcription factor binding (particularly pioneer factors) and chromatin remodelling enzymes. These factors can influence both nucleosome positioning (the location of a nucleosome along the DNA sequence) and occupancy (the frequency with which a specific site is incorporated into a nucleosome) (Bell *et al.*, 2011).

The nucleotide composition of a DNA sequence can influence how likely it is to be incorporated into a nucleosome. DNA must bend sharply around the core histones and therefore nucleosome formation is favoured by intrinsically flexible sequences and is disfavoured by more rigid sequences (such as stretches of deoxyadenosine on one strand of DNA, known as poly(dA:dT) tracts) (Struhl and Segal, 2013). This led to the hypothesis that nucleosome positioning and occupancy may be encoded within the genomic DNA sequence. Indeed, in some eukaryotes, such as the yeast species *Saccharomyces cerevisiae*, poly(dA:dT) tracts are highly enriched in nucleosome-depleted regions of gene promoters. A small number of *S. cerevisiae* promoters are nucleosome-depleted but lack poly(dA:dT) tracts, suggesting that chromatin is regulated at these sites by alternative mechanisms, such as the recruitment of chromatin remodellers (Struhl and Segal, 2013). However, in other eukaryotes, including humans, the role of DNA sequence features such as poly(dA:dT) tracts appears to be less important (Bell *et al.*, 2011). Even in yeast, correct nucleosome positioning, occupancy and spacing *in vitro* appear to be highly dependent on chromatin remodelling enzymes, which can override the intrinsic sequence preferences of nucleosomes (Zhang *et al.*, 2011).

Interestingly, the genes regulated by the *S. cerevisiae* promoters completely lacking poly(dA:dT) tracts were related to the stress response, a subset of genes that is likely to require more complex signal-dependent regulation (reviewed in Struhl and Segal,

2013). Similarly, chromatin accessibility is an important regulator of transcription factor binding to mammalian enhancers, which integrate various regulatory signals to control cell-type-specific gene expression. Several studies have demonstrated that enhancers are often enriched for DNA shape and sequence features that contribute to higher intrinsic nucleosome occupancy, thereby protecting against inappropriate transcription in the absence of cell-specific regulatory signals (Barozzi *et al.*, 2014; Tillo *et al.*, 2010). Accessible chromatin patterns at distal enhancers are highly cell-type-specific, as shown by ENCODE DNase hypersensitivity data from 125 cell and tissue types. Approximately one-third of the 2.9 million DNase hypersensitive sites (DHSs) identified in this study were specific to a single cell type, with the majority of DHSs located in distal intronic and intergenic regions. Conversely, gene promoters were typically highly accessible across many cell types (Thurman *et al.*, 2012).

An example of the regulatory role of chromatin accessibility in transcriptional regulation (discussed previously) is glucocorticoid receptor (GR), which relies on pre-determined patterns of chromatin accessibility for binding at 95% of hormone-induced sites (John *et al.*, 2011). Cell-specific chromatin accessibility is mediated by the action of co-operating transcription factors, such as the pioneer factor AP1 in murine mammary epithelial cells (Biddie *et al.*, 2011). Another example is oestrogen receptor (ER) which, like GR, demonstrates highly cell-type-specific binding. A study by Gertz *et al.* (Gertz *et al.*, 2013) analysed the features of ER binding sites that were shared between two ER-responsive cell lines compared to those that were specific to just one line. The main features of cell-specific binding sites were a lack of EREs (or weak EREs), co-occurrence with other transcription factor motifs and cell-type-specific accessible chromatin. This study suggests that it is the absence of strong ER motifs that enables chromatin to act as a regulator in this genomic context. Without additional regulatory inputs (in this case the action of co-operating transcription factors such as FOXA1, GATA3 and ETV4), ER is unable to bind to DNA at these sites to regulate gene expression. Conversely, ER binding sites that were common to both cell lines were characterised by high-affinity oestrogen response elements (EREs), facilitating strong ER binding regardless of the baseline level of chromatin accessibility.

Although intrinsically inhibitory to transcription factor binding, chromatin structure may become even more inhibitory through the actions of repressors such as the Polycomb group proteins. Polycomb group proteins form multi-sub-unit complexes known as Polycomb repressive complexes 1 and 2 (PRC1 and PRC2), which typically co-localise at genomic sites and contribute to the formation of repressive Polycomb chromatin

domains. PRC1 catalyses the ubiquitination of histone H2A at lysine 119 (H2AK119ub1), while PRC2 adds methyl groups to H3 at lysine 27 (H3K27me3). Sub-units of the Polycomb complexes can bind to these same histone modifications, contributing to the spreading and maintenance of Polycomb domains (Blackledge *et al.*, 2015). However, the exact mechanisms by which vertebrate Polycomb complexes act to repress transcription are not well understood. H2AK119ub has been shown to inhibit transcriptional elongation (Zhou *et al.*, 2008), although other studies suggest that Polycomb complexes are recruited to regulatory elements after transcriptional silencing has already been achieved (reviewed in Blackledge *et al.*, 2015). There is also little evidence that the histone modifications catalysed by the core sub-units directly affect chromatin structure by themselves (Blackledge *et al.*, 2015). A recent imaging study suggests that *Drosophila* Polycomb complexes contribute to dense chromatin packaging that increases with domain length and strong exclusion of nearby transcriptionally active chromatin regions (Boettiger *et al.*, 2016). However, the molecular mechanisms by which Polycomb proteins facilitate this unique chromatin structure remain unknown.

## **Histone Modifications**

Histone modifications can also influence chromatin structure, thereby regulating transcription factor accessibility. As discussed previously, there is an ongoing debate as to whether histone modifications are instructive or are defined by the transcriptional state. Regardless of their origin, it is clear that histone modifications can have direct structural effects on nucleosomes, which can affect transcription. Other regulatory effects of histone modifications include the recruitment and/or stabilisation of some transcription factors and the coupling of transcription to metabolic state. In all of these situations, however, it should be remembered that transcription factors themselves are likely integral to the processes that place these potentially regulatory modifications.

Recently, there has been renewed interest in the role of histone modifications in regulating nucleosome structure. Many new modifications have been identified in the globular domains of histone proteins. The positively-charged lateral surface of the globular domains is in contact with DNA, making 14 points of contact and more than 120 direct atomic interactions (Cosgrove *et al.*, 2004). Modifications that affect the lateral surface therefore have the potential to directly influence histone-DNA interactions. Acetylation, in particular, neutralises the positive charge of lysine and arginine residues, which may weaken the interaction with negatively-charged DNA. Several examples of lateral surface modifications have been characterised, including

H3K56, H3K64 and H3K122 acetylation, H3R42 asymmetric dimethylation and H3T118 phosphorylation (reviewed in Tessarz and Kouzarides, 2014; Tropberger and Schneider, 2013). H3K56 is located close to the DNA entry-exit region of the nucleosome and acetylation of this residue increases the rate of partial unwrapping of DNA from nucleosomes (“nucleosome breathing”), increasing access for regulatory factors (Neumann *et al.*, 2009). H3K56ac may also alter higher-order chromatin structure (Watanabe *et al.*, 2010). H3K64 and H3K122 are located close to the nucleosome dyad axis (approximate two-fold axis of symmetry located at the interface between the H3-H4 dimers). Histone-DNA interactions are strongest at the dyad axis (Hall *et al.*, 2009) and acetylation of these residues promotes nucleosome destabilisation and eviction. Both H3K64 and H3K122 acetylation have also been functionally linked to transcriptional activation *in vitro*, which, in the absence of any identified “reader” proteins, is believed to be mediated by direct structural effects (Di Cerbo *et al.*, 2014; Tropberger *et al.*, 2013). In addition, H3K64 and H3K122 acetylation is enriched at active promoters and a subset of active enhancers, consistent with a role in transcriptional activation (Pradeepa *et al.*, 2016). These examples demonstrate two distinct mechanisms by which histone globular domain modifications may regulate transcription factor accessibility.

Other histone modifications may have regulatory effects through co-transcriptional mechanisms such as histone exchange. The histone tail modification H3K36 methylation, for example, is enriched in the gene bodies of actively transcribed genes, although it has been associated with a repressive effect on transcription (Strahl *et al.*, 2002). The Set2 methyltransferase responsible for this modification is recruited to sites of active transcription by the phosphorylated CTD of Pol II (Krogan *et al.*, 2003). H3K36 methylation results in recruitment of the yeast Rpd3S histone deacetylase complex. In addition, H3K36 methylation inhibits H3 interaction with histone chaperone proteins and recruits the chromatin remodeller Isw1b, both of which prevent histone exchange and subsequent incorporation of pre-acetylated histones (Smolle *et al.*, 2012; Venkatesh *et al.*, 2012). Together these mechanisms function to maintain transcribed gene bodies in a hypoacetylated state in order to prevent the initiation of transcription from cryptic promoters.

Some histone modifications may also directly regulate transcription factors. FOXA1, for example, has been shown to recognise H3K4 mono- and di-methylation and reduced levels of H3K4me1/2 impair FOXA1 binding (Lupien *et al.*, 2008). Another example is the transcription factor AIRE, which binds specifically to non-methylated H3K4

(H3K4me0) through its PHD finger, facilitating the ectopic expression of tissue-specific antigens within the thymus (Org *et al.*, 2008). Once again, however, it is not completely clear whether these modifications actively guide transcription factor binding or are a result of transcriptional processes acting at these sites. Another study, for example, suggests that FOXA1 facilitates the deposition of H3K4me1/2 through recruitment of MLL3 (Jozwik *et al.*, 2016). A possible scenario is that both of these mechanisms operate, with transcription factor binding facilitating specific histone modifications, which then subsequently recruit or stabilise additional transcription factor binding to establish positive feedback loops.

Almost all histone modifying enzymes also have non-histone substrates, including transcription factors. Although this is not a direct effect of histone modifications themselves, the effects of these enzymes on transcription factor function can be significant. GATA1, for example, can be acetylated by the histone acetyltransferase CREBBP at several lysine residues near the DNA-binding zinc fingers. Bromodomain-containing protein 3 (BRD3) binds to acetylated GATA1 through one of its two bromodomains and this interaction is essential for the correct targeting of GATA1 to erythroid promoters (Gamsjaeger *et al.*, 2011; Lamonica *et al.*, 2011). Thus, transcription factors may be covalently modified just like histone proteins and these modifications may be “read” by co-factor domains.

Finally, histone modifications may regulate transcription by coupling it to the metabolic state of the cell. HAT-mediated acetylation, for example, requires acetyl-coenzyme A (acetyl-CoA) and is inhibited by CoA, while methylation requires S-adenosyl-methionine (SAM, produced from ATP and methionine) and is inhibited by S-adenosyl-L-homocysteine (SAH, formed by the demethylation of SAM). Similarly, the Sirtuin family of HDACs requires nicotinamide adenine dinucleotide (NAD<sup>+</sup>), lysine demethylases 1 and 2 require flavin adenine dinucleotide (FAD<sup>+</sup>) and the JumonjiC domain-containing demethylases require  $\alpha$ -ketoglutarate and Fe<sup>2+</sup> (Kinnaird *et al.*, 2016; van der Knaap and Verrijzer, 2016). Levels of these activating and inhibitory metabolites are influenced by various factors, including nutrient availability, caloric intake and the Circadian clock, and may provide a mechanism to couple metabolic cues with gene regulation (van der Knaap and Verrijzer, 2016). Studies in yeast, for example, demonstrate that the availability of acetyl-CoA determines the level of histone acetylation and gene expression (reviewed in van der Knaap and Verrijzer, 2016). The interconnections between metabolism, chromatin and transcription are just beginning to be explored and remain an exciting area of ongoing research.

## Histone Exchange and Variants

Histone proteins are not static components of nucleosomes but can be removed and replaced in a dynamic manner. The process of histone exchange is regulated by a number of factors, many of which have been previously discussed. Histone-DNA interactions may be disrupted by the action of ATP-dependent chromatin remodellers, for example, causing nucleosome sliding or eviction and facilitating histone exchange. Post-translational histone modifications can also promote or inhibit histone exchange (for example, the methylation of H3K36, discussed in the previous section). In addition, the process of transcription itself can stimulate histone exchange. One consequence of the dynamic turnover of histones is that canonical histones may be replaced by histone variants, with subsequent effects on chromatin structure and transcription factor accessibility (Venkatesh and Workman, 2015). Along with histone modifications, the incorporation of histone variants provides an additional mechanism for chromatin-mediated gene regulation, through direct effects on nucleosome structure and indirect effects as a result of differential protein interactions.

The canonical histones are each encoded by multiple genes that lack introns, are synthesised only during S-phase and are incorporated into DNA in a replication-dependent manner. In contrast, histone variants are encoded by unique genes, which contain introns (and may therefore undergo alternative splicing), and are synthesised and incorporated into DNA throughout the cell cycle (replication-independent). Human histone variants have been identified for all canonical histones except for H4. There are currently eight known human variants of H2A, six variants of H3 and two variants of H2B (Buschbeck and Hake, 2017). H3 and H2B variants generally differ from their core counterparts by only a few amino acids, whereas H2A variants contain more substantial alterations (Maze *et al.*, 2014).

The process of histone exchange is facilitated by histone chaperones and specific chromatin remodellers. Chaperones bind histone proteins and have diverse functions in histone storage, transport, nucleosome assembly and nucleosome disassembly. Most chaperone proteins are specific for either H3-H4 or H2A-H2B complexes and may have additional specificity for histone variants (Venkatesh and Workman, 2015). Chromatin remodellers also have direct roles in histone exchange. The Snf2 related CREBBP activator protein (SCRAP) and INO80 chromatin remodellers, for example, are required for the deposition and removal respectively of the histone variant H2A.Z (Buschbeck and Hake, 2017; Papamichos-Chronakis *et al.*, 2011).

Two histone variants that are temporally and spatially associated with transcription are H3.3 and H2A.Z. The H3 variant H3.3 differs from canonical H3 (H3.1 and H3.2) by only 4-5 amino acids. H3.3 interacts with the H3.3-specific chaperone HIRA, which facilitates the deposition of H3.3 in transcribed regions (gene bodies, promoters and enhancers). In contrast, canonical H3-H4 tetramers interact with histone chaperones such as anti-silencing function 1 (ASF1) and the chromatin assembly factor 1 (CAF1) complex (Buschbeck and Hake, 2017). H3.3 deposition in gene bodies requires transcriptional elongation and H3.3 was originally thought to function as a simple replacement for H3 in nucleosomes disrupted by the passage of RNA Pol II (Talbert and Henikoff, 2017). However, H3.3 may also play a more active role in transcription, as it has been shown to be required for gene expression in response to signals such as retinoic acid, interferon-gamma and heat shock (reviewed in Teves *et al.*, 2014). H3.3 does not significantly affect the structure of single nucleosomes, however it does inhibit higher-order chromatin compaction (Chen *et al.*, 2013b). H3.3 has also been shown to be targeted by a different chaperone to telomeres and heterochromatin, although the role of H3.3 at these locations is unclear (Teves *et al.*, 2014).

H2A.Z is also associated with transcription. Unlike H3 and H3.3, H2A.Z.1 and H2A.Z.2 have only 60% sequence similarity to canonical H2A. The deposition of H2A.Z is facilitated by the chromatin remodeller SRCAP and is enriched around transcription start sites (Teves *et al.*, 2014). H2A.Z may destabilise nucleosome structure, for example through altering the interface between the H2A.Z-H2B dimer and the H3-H4 tetramer, facilitating access to DNA by regulatory factors when transcription is induced (Venkatesh and Workman, 2015). The incorporation of both H2A.Z and H3.3 within the same nucleosome leads to a decrease in stability and the co-localisation of these two variants is common at DHSs including promoters and enhancers (Jin *et al.*, 2009). The dynamic exchange of H2A.Z-H2B dimers during transcription may also facilitate the passage of RNA Pol II during transcriptional elongation (Teves *et al.*, 2014). However, similarly to H3.3, the H2A.Z variant has also been shown to be associated with gene repression and heterochromatin, suggesting additional context-specific functions.

The incorporation of histone variants into nucleosomes can have direct and indirect effects on chromatin structure, which in turn influences the accessibility of DNA to regulatory factors. Direct effects include alterations in nucleosome stability and higher-order chromatin structures. Indirect effects may be mediated by differential interactions of variants with histone chaperones, histone modifying enzymes (which may catalyse or bind to variant-specific modifications) and chromatin remodelling complexes.

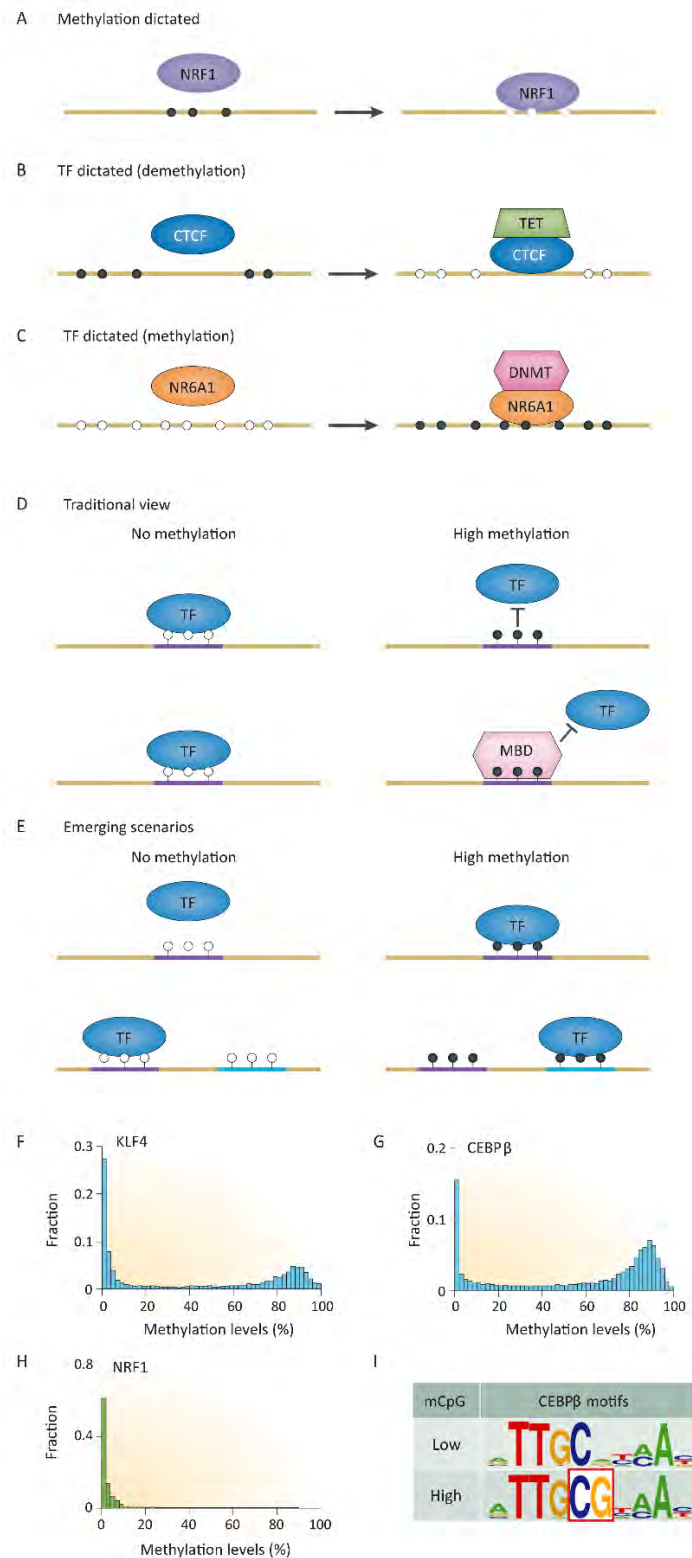
## DNA Methylation

Transcription factors can influence the DNA methylation landscape; however, their function may also be regulated by DNA methylation. The balance between these reciprocal effects may to some extent be determined by the properties of individual transcription factors (and/or their co-factors) - some transcription factors are strongly inhibited by DNA methylation, while others can recruit enzymatic machinery to alter methylation levels (Figure 1.13A-C). Furthermore, methylated CpGs may represent novel binding sites for some transcription factors, modifying their DNA binding specificity and potentially their regulatory function (Zhu *et al.*, 2016).

Methylation can affect, either positively or negatively, how transcription factors are able to interact with DNA. The addition of a hydrophobic methyl group to carbon 5 of cytosine, positioned in the DNA major groove, may alter the contacts between the transcription factor and DNA (base readout). In addition, the methyl group is bulky and may therefore alter the three-dimensional structure of the transcription factor binding site (shape readout). Each of these effects may either enhance or inhibit binding, depending on the DNA recognition mechanism used by the transcription factor (Dantas Machado *et al.*, 2015). Methyl-CpG binding proteins may also physically block transcription factor binding sites, as well as recruit additional co-factors to create a repressive chromatin structure, thereby indirectly inhibiting transcription factor binding (Zhu *et al.*, 2016).

Some transcription factors are strongly inhibited by DNA methylation. One example is nuclear respiratory factor 1 (Nrf1), which contains several CpGs in its consensus motif (Figure 1.13A) (Mathelier *et al.*, 2016). In a study of murine embryonic stem cells with genetic deletion of all three DNMTs, causing global DNA demethylation, more than 7,000 Nrf1 binding sites were increased compared to wild-type cells. These sites of increased binding were enriched in distal low-CpG regions and were associated with increased DNase hypersensitivity and increased expression of the nearest gene (Domcke *et al.*, 2015). Another transcription factor, CTCF, has also been reported to be inhibited by methylation of its core binding site. Methylation of the insulin like growth factor 2 / H19 enhancer, for example, controls the allele-specific expression (imprinting) of these genes by preventing the interaction of CTCF zinc finger 7 with DNA (Bell and Felsenfeld, 2000; Renda *et al.*, 2007). Interestingly, at the genome-wide level, methylation of CTCF binding sites does not appear to be a primary mechanism for regulation of CTCF binding, with the majority of CTCF binding sites remaining unchanged on global DNA demethylation; however, a small set of CTCF binding sites





**Figure 1.13: Interactions between transcription factors and methylated DNA**

**Figure reproduced (and caption modified) with permission from Nature Publishing Group (Zhu et al., 2016)**

(A) Binding of transcription factors to DNA may be inhibited by DNA methylation. The filled circles represent methylated DNA and the open circles represent unmethylated DNA.

(B) CTCF can initiate local DNA demethylation, presumably by recruitment of TET enzymes. (C) Conversely, the binding of NR6A1 increases DNA methylation by interacting with DNA methyltransferases. (D) Traditional view of interactions between transcription factors (TFs) and DNA. TFs bind to unmethylated DNA in regions of open chromatin. Methylation of CpG dinucleotides in the TF binding site may directly inhibit TF binding by affecting base or shape readout (see text). Alternatively, methyl-CpG binding-domain (MBD) proteins may bind to the methylated sequence and indirectly inhibit TF binding. MBD proteins may also recruit additional co-factors to create a repressive chromatin structure. (E) Emerging scenarios for interactions between TFs and DNA. DNA methylation may create a new binding site for TFs. TFs may also be able to recognise different sequences when DNA is unmethylated compared to when it is methylated. (F) Integration of DNA methylation and transcription factor ChIP-seq data for KLF4 in the H1 human embryonic stem cell line. The x axis shows average methylation level (percentage) of the CpG sites within a ChIP-seq peak and the y axis shows the fraction of peaks with a certain average methylation level. (G) Integration of DNA methylation and transcription factor ChIP-seq data for CEBPb. (H) Integration of DNA methylation and transcription factor ChIP-seq data for the control TF NRF1 (not known to interact with methylated DNA). (I) Motifs identified for CEBPb ChIP-seq peaks with low methylation (<60% average level) or high methylation (>80% average level). The methylated CpG site is outlined in red. CEBPb, CCAAT/enhancer binding protein beta; CTCF, CCCTC-binding factor; DNMT, DNA methyltransferase; KLF4, Kruppel-like factor 4; MBD, methyl-binding domain protein; mCpG, methylated CpG dinucleotide; NR6A1, nuclear receptor 6A1; NRF1, nuclear respiratory factor 1; TET, ten-eleven translocation enzyme; TF, transcription factor.

(around 6.6% of total) are affected, primarily reflecting a reactivation of CTCF sites specific to other tissue types (Maurano *et al.*, 2015). In addition, CTCF has been shown to initiate DNA demethylation at low-CpG regulatory regions (Figure 1.13B) (Feldmann *et al.*, 2013; Stadler *et al.*, 2011). Therefore, it appears that the effects of methylation on a transcription factor are context-dependent, although the additional factors involved in dictating this dependency are unknown. Furthermore, the same transcription factor can both regulate and be regulated by DNA methylation, indicating that these effects are not mutually exclusive.

It is becoming increasingly recognised that methylation may enhance the binding of some transcription factors, leading to a revision of the traditional inhibitory model (Figure 1.13D-E). Early studies identified zinc finger and BTB domain containing 33 (ZBTB33, also known as Kaiso), ZBTB4 and ZBTB38 as sequence-specific binders of methylated DNA. These proteins were found to interact with 5mC through their classical (C2H2) zinc finger domains and to act as transcriptional repressors in *in vitro* studies (Filion *et al.*, 2006; Prokhortchouk *et al.*, 2001). Multiple studies have since

attempted to identify methyl-CpG binding proteins, using methods such as mass spectrometry, protein microarrays, DNA microarrays, and chromatin immunoprecipitation followed by bisulphite sequencing (ChIP-BS-seq) (reviewed in Zhu *et al.*, 2016). A recent protein microarray study, for example, screened 1321 transcription factors and 210 co-factors for methylation-specific interactions with 154 DNA sequences (Hu *et al.*, 2013). The study identified 41 transcription factors that showed methyl-CpG-dependent binding, with the majority showing significant sequence specificity. The identified transcription factors belonged to diverse families, including the zinc-finger, Homeobox, basic helix-loop-helix and Forkhead box families. Kruppel-like factor 4 (KLF4) was one of the methyl-CpG-binding transcription factors identified and was shown to interact with distinct methylated and unmethylated sequences via two different DNA-binding domains (Hu *et al.*, 2013).

The correlation of genome-wide transcription factor ChIP-seq and whole-genome bisulphite sequencing from human embryonic stem (H1) cells has provided evidence of *in vivo* binding of candidate transcription factors to methylated DNA (Hu *et al.*, 2013; Zhu *et al.*, 2016). Zhu *et al.* (2016) analysed ChIP-seq data from H1 cells for 6 transcription factors identified as potential methyl-CpG binding proteins. The average binding site methylation levels for two of these transcription factors, CCAAT/enhancer binding protein beta (CEBPb) and KLF4, are shown in Figure 1.13F-G. A control transcription factor (NRF1), not known to interact with methyl-CpG, is shown in Figure 1.13H. A bimodal distribution was observed for 5 of the 6 candidate methyl-CpG binding proteins, indicating that a large fraction of their binding sites were located in highly methylated regions. CEBPb, for example, had an average methylation level of >80% for 6,675 (43%) of its binding peaks. Conversely, almost all NRF1 binding sites had low methylation levels. Motif analysis of low methylation peaks (<60% average methylation) and high methylation peaks (>80%) showed enrichment for distinct motifs, with the high methylation peaks containing a prominent methylated CpG dinucleotide (Figure 1.13I). For KLF4, reversal of the methylation status of the two distinct motifs inhibits the binding of KLF4 to both sequences (Hu *et al.*, 2013). Thus, for methyl-CpG-binding transcription factors, the methylated recognition sequences may be markedly different from the non-methylated recognition sequences and this may provide a dynamic mechanism to create or remove transcription factor binding sites. This has led to the description of 5mC as the “fifth nucleotide” (Zhu *et al.*, 2016).

The first methyl-CpG-binding transcription factors (for example, Kaiso) were shown to be associated with repressive functions *in vitro*. However, there are now several known

examples where binding of a transcription factor to methylated DNA can result in gene activation. These include KLF4, CCAAT/enhancer binding protein alpha (CEBPa) and regulatory factor X1 (RFX1) (Hu *et al.*, 2013; Niesen *et al.*, 2005; Rishi *et al.*, 2010). Cebpa, for example, was shown to bind to methylated sequences in low-CpG murine promoters to activate the expression of keratinocyte- or adipocyte-specific genes during differentiation. DNA methylation was essential for Cebpa binding at these sites, indicating a novel role for low-density CpG methylation in the creation of new transcription factor binding sites for tissue-specific gene activation (Rishi *et al.*, 2010).

Once transcription is initiated, DNA methylation may also regulate transcriptional elongation and gene splicing. Gene body methylation is associated with transcriptional activation (Dunham *et al.*, 2012). Recently, Dnmt3b was shown to be recruited to actively transcribed gene bodies by SetD2-mediated H3K36me3, where it functions to prevent aberrant transcription initiation from intragenic regions (Neri *et al.*, 2017). Exons are also more highly methylated than introns, suggesting a potential role of gene methylation in the regulation of splicing (Jones, 2012). In a direct example of this, CTCF has been shown to bind to an intragenic region within exon 5 of protein tyrosine phosphatase receptor type C (PTPRC, also known as CD45). Binding of CTCF to this region increases RNA Pol II pausing, promoting the inclusion of exon 5 in the transcript. Conversely, methylation of this intragenic site inhibits CTCF binding and promotes exclusion of the exon (Shukla *et al.*, 2011). This demonstrates a direct role for gene body methylation in the regulation of alternative splicing and it seems likely that more functions for this dynamic gene body modification will be discovered.

DNA methylation can have effects on transcription factor function that go far beyond inhibition of binding. However, there are still unanswered questions around how DNA methylation fits into the wider regulatory context. Firstly, as discussed previously, it is unclear if methylation is the initiating event in these effects or if methylation acts to reinforce transcriptional and chromatin states set in motion by other regulatory mechanisms. This may also apply to contexts beyond gene repression. The binding of Cebpa to low-CpG methylated promoters, for example, does not appear to be driven by changes in methylation that occur during differentiation, even though methylation is required for binding at these sites to occur (Rishi *et al.*, 2010). Secondly, the effects of methylation on a transcription factor can be context-dependent, for example in the case of CTCF. It is unclear, however, what additional regulatory factors interact with methylation to cause these differences. Finally, the role of dynamic DNA methylation in low-CpG regulatory regions such as enhancers is still being explored. Intergenic

regions and gene bodies are the most variably methylated regions in the human genome and lineage-specific differences in enhancer methylation have been observed (Dunham *et al.*, 2012; Schmidl *et al.*, 2009). KLF4 was also shown to bind to methylated enhancer regions, hinting at a role for methyl-CpG-binding transcription factors in enhancer regulation (Zhu *et al.*, 2016). Additional research into these areas may further illuminate the regulatory intricacies of DNA methylation beyond gene silencing.

### 1.3 Breast Cancer

Breast cancer is the second-most common cancer worldwide, with approximately 1.67 million new cases diagnosed in 2012 (Ferlay J, 2013). In Australia, it is estimated that almost 18,000 people will be diagnosed with breast cancer in 2017 (144 males and 17,586 females). Breast cancer is the most frequently diagnosed cancer in women, representing 28% of all cancer diagnoses and carrying an estimated 1 in 8 risk of diagnosis before age 85 (Australian Institute of Health and Welfare, 2017).

The overall 5-year survival rate for breast cancer in 2017 is 90%. Significant improvements in survival have occurred over the last 25 years due to the introduction of population-based screening (for example, BreastScreen Australia in 1991), as well as advances in treatments (for example, the development of ERBB2-targeting antibody therapies) (Australian Institute of Health and Welfare, 2017). Mammographic screening aims to reduce mortality through the early detection of disease, although interpretation of efficacy can be complex. As an example, screening results in an apparent increase in breast cancer incidence (including the detection of some cancers that would never have presented clinically), as well as an increase in time from diagnosis to death regardless of screening efficacy. Recent meta-analyses that incorporate these complexities estimate that screening in women aged 55-79 reduces breast cancer mortality by approximately 20% (Marmot *et al.*, 2013).

Despite these advances, however, approximately 3,000 Australians will die from breast cancer in 2017, making it the second-most common cause of cancer death in females after lung cancer (Australian Institute of Health and Welfare, 2017).

#### **Two Clinical Challenges: Endocrine-resistant Disease and “Triple-Negative” Breast Cancers**

More than 70% of breast cancers express oestrogen receptor (ER) and the introduction of therapies targeting ER signalling has led to dramatic improvements in survival. Tamoxifen treatment, for example, reduces the 10 year recurrence risk by almost half (Early Breast Cancer Trialists' Collaborative Group, 1998). However, a number of patients will present with *de novo* resistance to endocrine therapies, and another 20-30% of patients will develop recurrent disease, often many years after the initial diagnosis. As a result, ER-positive breast cancers are responsible for more patient deaths overall than ER-negative (including “triple-negative” and HER2-positive) breast cancers (Clarke *et al.*, 2015; Musgrove and Sutherland, 2009).

Another challenge is presented by triple-negative breast cancers (TNBCs), which do not express ER or progesterone receptor (PR), and which lack overexpression of the erb-b2 receptor tyrosine kinase 2 (ERBB2, also known as HER2). Unlike ER-positive disease (which can be targeted with endocrine therapies) and HER2-positive disease (which has benefited from the recent implementation of HER2-targeting therapies), there is a lack of targeted molecular therapies for TNBC. Furthermore, TNBCs are often more aggressive, carry a higher risk of relapse in the first 5 years after treatment, and have poorer overall survival compared to other breast cancer types. As a group, TNBCs are remarkably diverse at the molecular level, adding to the challenge of identifying and implementing targeted treatments (Bianchini *et al.*, 2016).

In both of the above, there is therefore a need for an improved understanding of the molecular drivers of these breast cancers, which will facilitate the development of targeted treatments. In addition, there is a need for novel biomarkers that can help predict the response of ER-positive tumours to endocrine therapy.

### **Clinical and Molecular Subtypes**

Breast cancer is a heterogeneous disease on many levels. Pathologists have long recognised that breast cancers can have a diverse range of histological appearances. The most common histological “type” is invasive ductal carcinoma (IDC) not otherwise specified (50-80%), with the remaining cases comprising a myriad of phenotypes including lobular (5-15%), medullary, tubular, cribriform, apocrine, metaplastic, and neuroendocrine carcinomas (Weigelt *et al.*, 2010). Clinically, individual patients exhibit different patterns of disease progression, sites of distant metastasis, and responses to treatment. Immunohistochemistry demonstrates distinct patterns of expression of hormone receptors and proliferation markers (Harbeck and Gnant, 2016; Weigelt *et al.*, 2010). Most recently, the molecular basis for breast cancer heterogeneity has begun to be unravelled through the use of microarray and sequencing technologies.

In 2000, a gene expression micorarray analysis of 42 normal and breast cancer samples (mostly invasive ductal carcinomas) led to the first classification of the molecular or “intrinsic” subtypes of breast cancer. These were derived from unsupervised hierarchical clustering of gene expression data, and were termed luminal A, luminal B, HER2-enriched, basal-like, and normal-like (Perou *et al.*, 2000; Sorlie *et al.*, 2001). An additional subtype known as claudin-low was identified in 2007 (Herschkowitz *et al.*, 2007). One of the most fundamental distinctions in breast cancer is the expression of ER, and this distinction is reflected in the intrinsic subtypes. The

general features of the intrinsic breast cancer subtypes are summarised in Table 1.4.

The luminal subtypes are characterised by the expression of genes that resemble the luminal epithelial cells of the breast, including ER, progesterone receptor (PR), luminal cytokeratins 8 and 18, and genes associated with ER activation. Expression of the transcription factors FOXA1, GATA3, MYB, and X-box binding protein 1 (XBP1), in addition to ER, are a core feature of luminal breast cancers. Luminal A tumours, in comparison to luminal B, tend to have higher expression of ER and ER-regulated genes, lower proliferation, lower frequency of p53 mutations, and low expression of the HER2 gene cluster (which can be low or high in Luminal B) (Perou and Borresen-Dale, 2011; Perou *et al.*, 2000).

The gene expression pattern of normal-like tumours closely resembles that of the normal breast. This is a diverse subgroup and includes tumours with high stromal and immune cell content, normal epithelial cell contamination and low tumour cell content (Perou and Borresen-Dale, 2011). It is unclear whether this represents a truly unique subtype or is an artefact of normal breast tissue within the samples.

The HER2-enriched, basal-like and claudin-low subtypes are generally ER-negative. The HER2-enriched subtype is characterised by increased expression of HER2-regulated genes, frequently accompanied by amplification and/or increased expression of the *HER2* gene (also known as *ERBB2*) and other genes located near the HER2 locus. However, approximately 30% of HER2-enriched tumours do not have clinical HER2 over-expression; these tumours may have activation of downstream elements of the HER2 signalling pathway. HER2-enriched tumours are also characterised by high expression of proliferation-associated genes, intermediate expression of luminal genes, and low expression of basal genes. A subset of clinically HER2-positive breast cancers also express luminal genes and are ER-positive, therefore placing them in the Luminal B subgroup. HER2-enriched tumours contain the highest number of mutations of any subtype and around 40% have mutations in p53 (Perou and Borresen-Dale, 2011; Prat *et al.*, 2015).

The basal-like subtype is frequently referred to as “triple-negative”, as in most cases (75%) these tumours lack expression of ER and PR, and do not overexpress HER2. Basal-like tumours have high expression of the basal gene cluster, which includes basal epithelial cytokeratins (5, 16, 14, 17), the receptor tyrosine kinases KIT and EGFR, vimentin, and P-cadherin. Other characteristic features of this subtype are the very low expression of luminal genes, relatively low expression of the HER2 gene



cluster, high expression of proliferation-associated genes, and a high occurrence of p53 mutations. Breast cancers in individuals with inherited BRCA1 mutations are most commonly of the basal-like subtype, although most basal-like breast cancers are sporadic (Perou and Borresen-Dale, 2011). On the basis of recent multi-platform, pan-cancer studies, basal-like breast cancer has been proposed to be a distinct molecular entity, showing more similarity to squamous cell lung cancer than the luminal subtypes of breast cancer (reviewed in Prat *et al.*, 2015).

The third main type of ER-negative breast cancer is the claudin-low subtype. These tumours have low expression of genes involved in tight junctions and cell-cell adhesion (for example, claudin 3, 4, 7, and E-cadherin) and high expression of mesenchymal genes such as vimentin, *SNAI1*, *SNAI2* and *TWIST1*. The claudin-low subtype also features high expression of immune-related genes, which may be expressed by the tumour cells or infiltrating immune cells (Perou and Borresen-Dale, 2011; Prat *et al.*, 2010). Gene expression studies of flow cytometry-sorted mammary epithelial cells have demonstrated that the expression profile of claudin-low cancer cells closely resembles that of mammary stem cells (defined as CD49f+/EpCAM-) (Lim *et al.*, 2009).

The intrinsic subtypes are associated with differences in response to treatments and survival. Basal-like and HER2-enriched tumours show the best response to chemotherapy, while luminal tumours are relatively chemoresistant. In terms of endocrine therapy, luminal A tumours are usually endocrine sensitive, while luminal B tumours are more variable in their response. Of all the subtypes, luminal A tumours are associated with the best prognosis in multiple studies (Prat *et al.*, 2015).

In 2009, the Prediction Analysis of Microarray (PAM50) was developed, allowing the prediction of intrinsic subtype based on the expression of 50 genes. PAM50 was shown to be an independent predictor of prognosis (Parker *et al.*, 2009). Clinically, however, the assessment of intrinsic subtype using tools such as PAM50 is not routine, and immunohistochemistry (IHC) and fluorescence *in situ* hybridisation (FISH) markers are frequently employed. These markers are the hormone receptors ER and PR (collectively HR) and HER2, leading to four primary classifications: HR+/HER2-, HR+/HER2+, HR-/HER2+, HR-/HER2-. Each of these classifications encompasses several intrinsic subtypes. The HR-/HER2- ("triple-negative") group, for example, includes basal-like (49%), claudin-low (30%), HER2-enriched (9%), luminal B (6%), and luminal A (5%) tumours (Prat and Perou, 2011). Proponents of molecular subtyping argue that the currently used markers do not adequately reflect the heterogeneity of breast cancer and that a detailed molecular understanding of tumours

will allow better prediction of prognosis and the use of more targeted therapies (Prat *et al.*, 2015). One example is that luminal A tumours that are clinically HER2+ have a similar outcome to luminal A tumours that are clinically HER2-; this subgroup of clinically HER2+ patients may therefore be suitable for less intensive chemotherapy than, for example, clinically HER2+ tumours that fall into the HER2-enriched or luminal B subtypes (Prat *et al.*, 2014).

The molecular subtyping of breast cancer continues to evolve, driven by new technologies and decreasing costs. In 2011, Lehmann *et al.* identified six distinct subtypes of triple-negative breast cancer based on gene expression profiles (basal-like (BL1, BL2), immunomodulatory (IM), mesenchymal (M), mesenchymal stem-like (MSL), and luminal androgen receptor (LAR) subtypes) (Lehmann *et al.*, 2011). Uniquely activated pathways were identified in each subtype, providing potential therapeutic targets, such as DNA repair in BL1 tumours, immune signalling in IM tumours, and AR in LAR tumours. Another study of almost 2,000 breast cancer samples, using microarray combined with genomic copy number analysis, identified 10 integrative clusters of breast cancer (Curtis *et al.*, 2012). A recent study of 127 invasive lobular carcinomas indicated that lobular carcinoma may be a distinct molecular, as well as histological, subtype (Ciriello *et al.*, 2015). Studies of many thousands of tumour samples using multiple platforms are also being performed by projects such as The Cancer Genome Atlas, providing molecular insights into cancer on a vast scale.

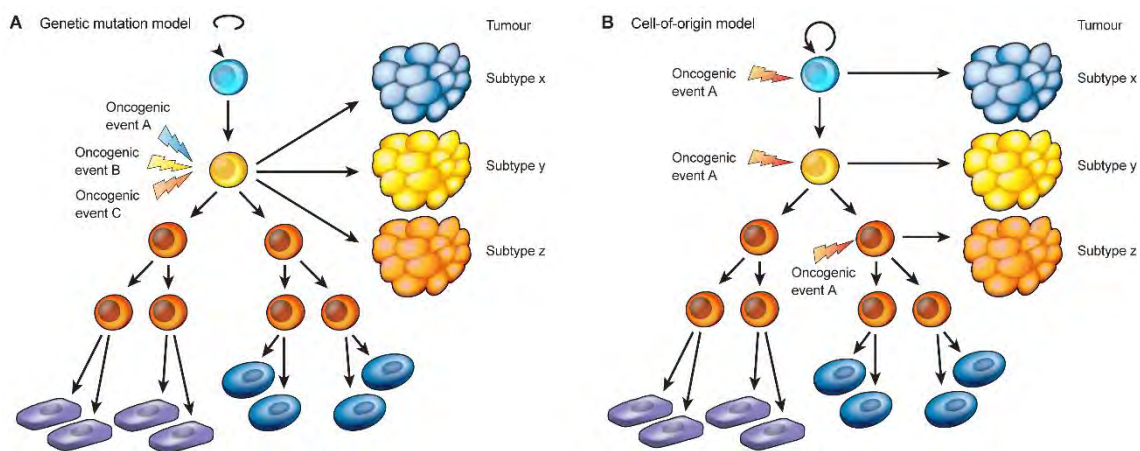
**Table 1.4: The intrinsic subtypes of breast cancer**

Subtype	Pathological features	Molecular features	Clinical features and treatment
Luminal A	ER+ PR+ HER2- Low Ki67 (proliferative marker)	High expression of luminal gene cluster (including ER, FOXA1, GATA3, XBP1, MYB) Low expression of proliferative cluster Low rate of p53 mutations	30% of breast cancers Endocrine sensitive Chemoresistant Associated with good prognosis Treatment: endocrine therapy (chemotherapy for selected patients)
Luminal B	ER+ PR+/- HER2- High Ki67, higher grade (vs A) Around 20% are HER2+	High expression of luminal gene cluster Lower expression of ER-related genes compared to luminal A (including PR, FOXA1) Higher expression of proliferative gene cluster compared to luminal A Higher rate of p53 mutations compared to luminal A	20% of breast cancers Less endocrine sensitive than Luminal A Mostly chemoresistant Anti-HER2 therapies may be effective in HER2+ cases Associated with poor prognosis Treatment: endocrine therapy +/- chemotherapy +/- anti-HER2 therapy
HER2-enriched	ER+/- PR +/- HER2+ Around 30% are HER2-	High expression of the HER2 gene cluster Most have HER2 amplification High expression of proliferative gene cluster Intermediate expression of luminal genes Low expression of basal genes	10-20% of breast cancers Good response to chemotherapy ER- cases have best response to chemotherapy but also have higher rate of relapse and poorer overall survival Associated with poor prognosis (prior to anti-HER2 therapies) Treatment: chemotherapy + anti-HER2 therapy
Basal-like	ER- PR- HER2- (75%) EGFR and CK5/6 can be used as IHC markers 10% are ER/PR+ 10-15% are HER2+	High expression of the basal gene cluster Low expression of luminal and HER2 gene clusters Low expression of luminal cytokeratins High rate of p53 mutations (80%) Common in patients with inherited BRCA1 mutation	10-20% of breast cancers More common in younger patients and African-Americans Good response to chemotherapy Triple-negative cases have best response to chemotherapy but also have higher rate of relapse and poorer overall survival ("triple-negative paradox") Associated with poor prognosis Treatment: chemotherapy
Claudin-low	ER- PR- HER2- (70%) Mostly high-grade Can have metaplastic and medullary differentiation 15-25% are ER/PR+	Low expression of cell adhesion proteins (claudin 3, 4, 7, E-cadherin) High expression of mesenchymal and extracellular matrix genes (vimentin, SNAI1, SNAI2, TWIST1) High expression of immune genes Low expression of luminal and HER2 gene clusters Low expression of proliferative gene cluster (but higher than luminal A or normal-like tumours) Inconsistent expression of basal cluster	Response to chemotherapy is intermediate between basal-like and luminal Associated with poor prognosis (similar to basal-like, luminal B, HER2-enriched) Treatment: chemotherapy

References: (Perou and Borresen-Dale, 2011; Perou *et al.*, 2000; Prat *et al.*, 2010; Prat and Perou, 2011; Prat *et al.*, 2015; Toss and Cristofanilli, 2015)

## Mammary Development and Breast Cancer Subtypes

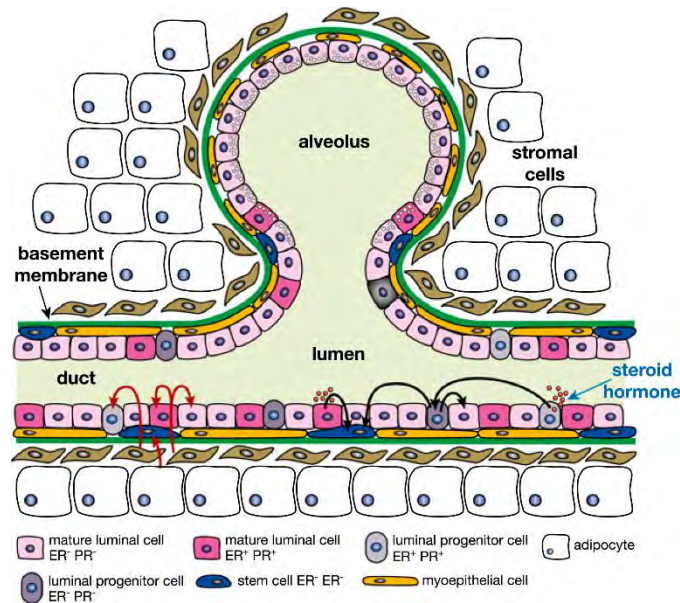
The source of heterogeneity in breast cancer is highly debated and two main models have been proposed (reviewed in Skibinski and Kuperwasser, 2015; Visvader, 2011). Firstly, different oncogenic events within the same target cell may result in distinct tumour phenotypes (genetic mutation model, Figure 1.14A). Alternatively, different tumour subtypes may arise from distinct cell types that exist within the organ or tissue (cell of origin model, Figure 1.14B). Another possibility is that a combination of these events may occur.



**Figure 1.14: Two models of intertumoural heterogeneity**

*Figure reproduced with permission from Nature Publishing Group (Visvader, 2011)*

Comparison of gene expression profiles from mammary epithelial cell subpopulations has revealed similarities with breast cancer subtypes, providing a framework in which to consider the above models. The normal mammary epithelium is composed of two layers of cells - an inner luminal layer and an outer, contractile layer of basal or myoepithelial cells. The epithelial cells form a series of branching ducts and lobules (containing the milk-producing cells), embedded within a stromal network (Figure 1.15). With every pregnancy, the mammary epithelium undergoes extensive cycles of proliferation and regression (Skibinski and Kuperwasser, 2015).



**Figure 1.15: The cellular organisation of the normal mammary gland**

**Figure adapted from (Visvader and Stingl, 2014)**

The normal mammary epithelium is composed of two layers of cells - an inner luminal layer and an outer, contractile layer of basal or myoepithelial cells. The epithelial cells form a series of branching ducts and lobules (containing the milk-producing alveolar cells), embedded within a stromal network containing fibroblasts and adipocytes. The mature luminal cells can be ER/PR<sup>-</sup> (light pink) or ER/PR<sup>+</sup> (dark pink). Luminal progenitor cells also exist within the ductal network. Oestrogen or progesterone (red circles in lumen) activate ER<sup>+</sup> epithelial cells, which secrete paracrine factors that activate other cell types. Mammary stem cells are located in the basal cell compartment. There is extensive signalling between different types of epithelial cells, as well as between epithelial and stromal cells (indicated by the red and black arrows).

Underlying this immense regenerative capacity is the mammary epithelial cell hierarchy. The mammary stem cell is defined by its unique ability to generate an entire functional mammary gland in murine transplant experiments (Shackleton *et al.*, 2006; Stingl *et al.*, 2006). This elusive cell accounts for about 1 in 50 cells within the basal epithelial population (Visvader and Stingl, 2014). Many details of the differentiation model remain controversial; it is unclear, for example, whether the mammary stem cell is bipotent (giving rise to both the luminal and basal lineages) or whether unique unipotent stem cells exist. Cells representing various differentiation states, defined by flow cytometry cell surface markers, can be identified in the mouse mammary gland, with similar populations identified in humans (Lim *et al.*, 2009). In general, the mammary stem cell(s) gives rise to luminal and basal progenitor cells, which then differentiate into mature luminal and myoepithelial cells. There are two types of mature

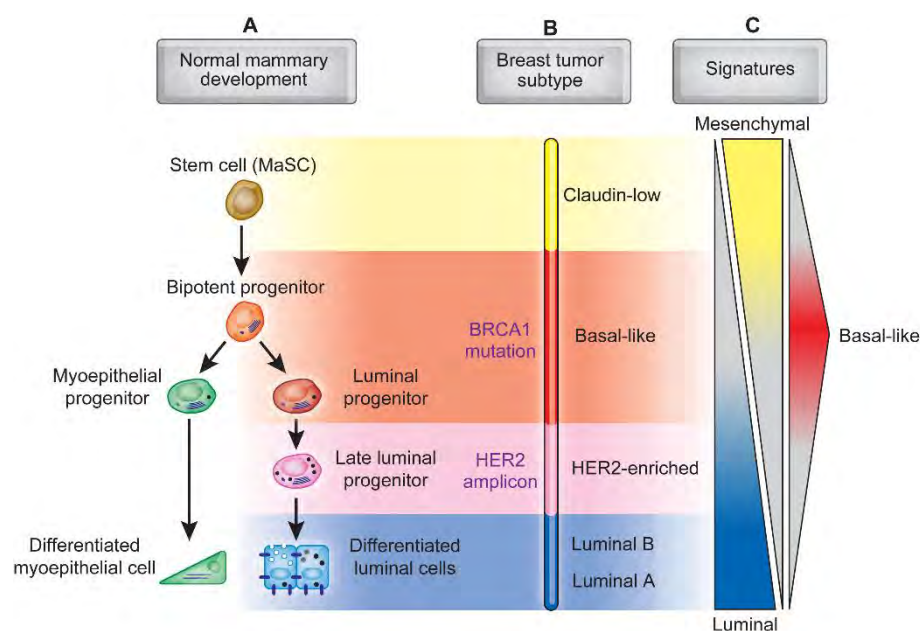
luminal cells, the ductal cells (or hormone-sensory cells, ER/PR-positive, ELF5-low) and the alveolar secretory cells (ER/PR-negative, ELF5-high) (reviewed in Visvader and Stingl, 2014).

Support for the “cell of origin” breast cancer model comes from the discovery that the gene expression signatures of breast cancer subtypes resemble the signatures of different types of mammary epithelial cells. The gene expression signature of claudin-low tumours, for example, is similar to the mammary stem cell, while luminal tumours express many of the same genes as mature ER-positive luminal cells. Interestingly, basal-like breast cancer cells most closely resemble the luminal progenitor cells (Lim *et al.*, 2009; Prat *et al.*, 2010). Human BRCA1 mutation carriers (who have a lifetime risk of developing basal-like breast cancer of >80%) demonstrate abnormal expansion and differentiation of the luminal progenitor cell population, providing further evidence that the luminal progenitor is the likely cell-of-origin in basal-like breast cancer (Lim *et al.*, 2009). In addition, deletion of *Brca1* in mouse basal cells results in tumours that do not resemble any of the common forms of mammary adenocarcinoma, arguing against the basal or mammary stem cell as the cell-of-origin in basal-like breast cancer (Molyneux *et al.*, 2010).

Studies have also revealed that there is a strong association between certain types of oncogenic events and different breast cancer subtypes. Examples include mutations in *GATA3*, *FOXA1*, *MAP3K1*, and *PI3K* in luminal A and B tumours, and the very high rate of p53 mutations (80%) in basal-like tumours. In addition, mouse models of breast cancer indicate that distinct oncogenic events in luminal progenitor cells can recapitulate much of the heterogeneity of the breast cancer subtypes (Melchor *et al.*, 2014). This suggests that the luminal progenitor cell could be a common cell-of-origin for multiple breast cancer subtypes, including basal-like and luminal A/B, with the nature of the oncogenic event determining the tumour phenotype (genetic mutation model). This is hypothesised to result from the alteration of cell fate pathways by oncogenic events (reviewed in Gross *et al.*, 2016; Skibinski and Kuperwasser, 2015).

The diagram in Figure 1.16 illustrates the similarities in the gene expression profiles that have been observed in mammary epithelial cells and breast cancer subtypes. As discussed above, one hypothesis is that each subtype arises from a distinct cell type, for example, claudin-low from the mammary stem cell, basal-like from the luminal progenitor, and luminal breast cancers from more differentiated cells. The cell-of-origin for the HER2-enriched subtype in this model is less clear, however the gene expression signature is consistent with a cell intermediate between the luminal

progenitor and mature luminal cell. Alternatively, breast cancer subtypes may arise from a common precursor cell (such as the mammary stem cell or luminal progenitor cell), subsequently undergoing altered differentiation programs depending on the underlying oncogenic event (Gross *et al.*, 2016; Prat and Perou, 2009). It is also possible that some combination of these events occurs. As an example, luminal progenitor cells are likely to be a heterogeneous population, containing cells at various points along a differentiation spectrum; some of these cells may be progressing towards an ER-positive hormone-sensing cell fate, while others are likely on the path to becoming ER-negative/ELF5-high secretory cells. The differentiation state of the luminal progenitor cell along this continuum, in combination with a particular oncogenic event, may therefore both contribute to the subtype of breast cancer that arises. Although the exact mechanisms are not completely understood, it is evident that normal development and breast cancer are intimately linked. An improved understanding of how these differentiation pathways are altered in cancer will provide insights into breast cancer development, progression, and therapeutic strategies.



**Figure 1.16: Model of the human mammary epithelial cell hierarchy linked to cancer subtype**

**Figure reproduced (and caption modified) with permission from Nature Publishing Group (Prat and Perou, 2009)**

(A) A simple representation of the normal mammary epithelial cell hierarchy.

(B) Comparison of gene expression profiles from mammary epithelial cell subpopulations

has revealed similarities with breast cancer intrinsic subtypes. Breast cancer subtypes may arise from epithelial cells at different points along the differentiation hierarchy. Alternatively, oncogenic events in a common cell type (for example, the mammary stem cell or luminal progenitor cell) may drive distinct differentiation pathways, giving rise to the various molecular subtypes. (C) The defining expression patterns of luminal, mesenchymal (claudin-low), and basal-like breast cancer cells. These molecular patterns may be best represented as gradients, rather than discrete on/off expression.



## 1.4 Transcription Factors in Cancer

Transcriptional dysregulation occurs commonly in cancer. Indeed, some of the earliest identified retroviral oncogenes (for example, *v-Myc* and *v-Jun*) were subsequently found to encode transcription factors (Vogt, 2012). In cancer, multiple molecular mechanisms can lead to aberrant transcription factor activity, driving gene expression programs that promote tumour initiation and progression. Part 1.4 of this chapter will focus on the ETS family of transcription factors, and the molecular mechanisms by which ETS factors can become abnormally activated or inactivated in cancer. This will be followed by a discussion of the known roles of the ETS transcription factor ELF5 in both normal development and cancer.

### The ETS Transcription Factor Family

ELF5 belongs to the evolutionarily conserved E26 transforming sequence (ETS) family of transcription factors. There are 28 known ETS factors in humans, which regulate fundamental cellular processes such as proliferation, differentiation, apoptosis and migration (Oikawa and Yamada, 2003). All ETS factors contain an ETS DNA-binding domain (approximately 85 amino acids), which recognises a core GGAA/T motif. The ETS family can be divided into 12 sub-families on the basis of sequence homology within the ETS domain. A subset of ETS factors also contain a second conserved domain known as the Pointed (PNT) domain (Table 1.5).

The ETS domain is a variant of the winged helix-turn-helix (wHTH) motif and contains three  $\alpha$ -helices and four  $\beta$ -sheets. Two arginine residues in helix 3 (part of the HTH motif) make contact with the GGAA/T DNA sequence and underpin the DNA-binding ability of ETS factors (Bosselut *et al.*, 1993; Findlay *et al.*, 2013). Interactions between the nucleotides surrounding the core ETS motif and distinct amino acid residues of the ETS factor DNA-binding domain also contribute to binding; this is believed to occur through an indirect mechanism (DNA backbone “shape readout”), as structural studies show that there is no direct contact outside the core DNA motif (reviewed in Hollenhorst *et al.*, 2011). Accordingly, four main classes of ETS DNA-binding specificities have been identified, which correlate well with ETS factor binding sites identified by ChIP-seq (Wei *et al.*, 2010). SPDEF, the sole member of class 4, is the only ETS factor to show a preference for the GGAT (rather than GGAA) core motif. As a class 2a ETS factor, the preferred ELF5 binding sequence *in vitro* is CCCGGAAGT, although multiple factors are likely to influence this preference *in vivo*. The function of the ETS domain, however, is not limited to DNA-binding; it is also an important site for


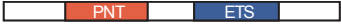










interactions with other transcription factors and co-factors (Sharrocks, 2001). The co-operative interactions of ETS1 with other transcriptional regulators such as PAX5 and RUNX1 (discussed previously), for example, all occur via the ETS domain.

Eleven ETS family members including ELF5 also contain a Pointed (PNT) domain (65-85 amino acids). Structurally, the PNT domain forms 4-5 compact  $\alpha$ -helices and is related to the Sterile Alpha Motif (SAM) domain, present in numerous eukaryotic proteins (Mackereth *et al.*, 2004). The Pointed domain has been linked to various function in ETS factors, including homo-oligomerisation (for example, ETV6, also known as TEL) and hetero-dimerisation (for example, between ETV6 and ETV7). The Pointed domain may also facilitate interactions with transcriptional co-factors or signalling proteins; examples include the interaction of the ETS1 Pointed domain with the kinase MAPK1, enhancing ETS1 transcriptional activity, and the interaction of ETV6 with the co-repressor SIN3A (reviewed in Garrett-Sinha, 2013; Hollenhorst *et al.*, 2011). Removal of the PNT domain in various ETS factor isoforms provides an additional level of regulatory control. While these findings point to a general role in protein-protein interactions, the specific functions of the PNT domain in several ETS factors, including ELF5, remain elusive.

**Table 1.5: The ETS transcription factor family**  
(next page)

The human ETS transcription factor family contains 28 members, which can be divided into 12 sub-families (column 1) on the basis of sequence homology within the DNA-binding ETS domain. In addition, four main classes of DNA-binding specificity have been identified (column 4); ETV3L was not included in this analysis (Wei *et al.*, 2010). A subset of ETS factors also contain a Pointed (PNT) domain, which closely resembles the Sterile Alpha Motif (SAM) domain. The general domain structure of the members of each sub-family is shown in column 5 (Sizemore *et al.*, 2017). Asterisk (\*) indicates that FEV, unlike other members of the ERG sub-family, does not contain a Pointed domain. All gene names and symbols are from the HUGO Gene Nomenclature Committee (HGNC) database, accessed September 2017 (Gray *et al.*, 2015), with several commonly used alternative gene names shown in parentheses. Sub-family abbreviations (largely based on original protein names) include: ERG, ETS-related gene; ESE, epithelium-specific ETS factor; ETS, E26 transforming sequence; PEA3, polyomavirus enhancer activator 3; SPI, spleen focus forming virus proviral integration oncogene; TCF, ternary complex factor; TEL, translocation, Ets, leukaemia.

**Table 1.5: The ETS transcription factor family**

Sub-family	Symbol	HGNC Gene Name	DNA-binding Class <sup>a</sup>	General Domain Structure
<b>ETS</b>	ETS1	ETS proto-oncogene 1, transcription factor	1	
	ETS2	ETS proto-oncogene 2, transcription factor	1	
<b>ERG</b>	ERG	ERG, ETS transcription factor	1	
	FLI1	Fli-1 proto-oncogene, ETS transcription factor	1	
	FEV*	FEV, ETS transcription factor	1	
<b>GABPA</b>	GABPA	GA binding protein transcription factor alpha subunit	1	
<b>ESE</b>	ELF3	E74 like ETS transcription factor 3	2a	
	ELF5	E74 like ETS transcription factor 5 (ESE2)	2a	
	EHF	ETS homologous factor	2a	
<b>TEL</b>	ETV6	ETS variant 6 (TEL)	2b	
	ETV7	ETS variant 7	2b	
<b>SPDEF</b>	SPDEF	SAM pointed domain containing ETS transcription factor	4	
<b>ELF</b>	ELF1	E74 like ETS transcription factor 1	2a	
	ELF2	E74 like ETS transcription factor 2	2a	
	ELF4	E74 like ETS transcription factor 4	2a	
<b>ERF</b>	ERF	ETS2 repressor factor	1	
	ETV3	ETS variant 3	1	
	ETV3L	ETS variant 3 like	NA	
<b>PEA3</b>	ETV1	ETS variant 1	1	
	ETV4	ETS variant 4 (PEA3)	1	
	ETV5	ETS variant 5	1	
<b>ETV2</b>	ETV2	ETS variant 2	1	
<b>SPI</b>	SPI1	Spi-1 proto-oncogene (PU.1)	3	
	SPIB	Spi-B transcription factor	3	
	SPIC	Spi-C transcription factor	3	
<b>TCF</b>	ELK1	ELK1, ETS transcription factor	1	
	ELK3	ELK3, ETS transcription factor	1	
	ELK4	ELK4, ETS transcription factor	1	

## ETS Factor Specificity

Many ETS factors show a degree of overlap in their genomic binding sites. In general, the sites bound by multiple ETS factors tend to contain high-affinity ETS consensus sequences, are located in close proximity to transcription start sites (20-40 bp) and show histone modifications consistent with active promoters. The genes regulated by these genomic binding events are also more likely to be highly expressed in multiple cell types, indicating a level of redundancy in ETS factor regulation of ubiquitously expressed “housekeeping” genes (reviewed in Hollenhorst *et al.*, 2011).

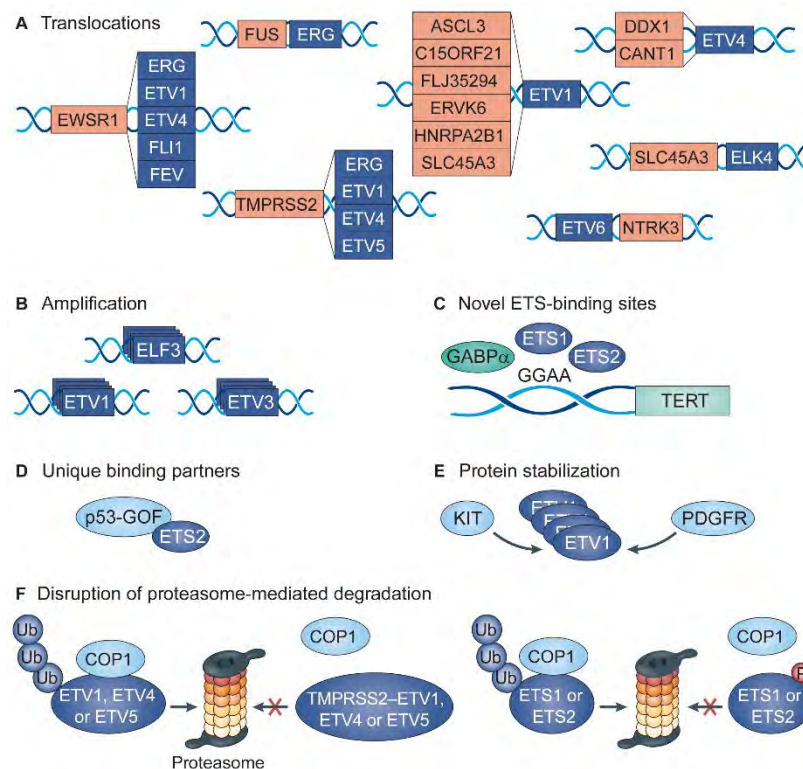
Despite the overall similarity in their DNA recognition sequences and extensive co-expression, however, ETS factors are also able to function as highly specific regulators of gene expression. This is achieved through many of the mechanisms described in the previous sections, including co-operativity with other transcription factors, interactions with co-factors, post-translational modifications, and auto-inhibition. Co-operative interactions or post-translational modifications may enable ETS factors to bind to “non-consensus” sites *in vivo*, with increasing evidence indicating that lower-affinity sites are important for precise and specific regulation of gene expression (Crocker *et al.*, 2016). Although structural studies are limited, the regions outside the main ETS and PNT domains appear to have minimal secondary or tertiary structure, which may provide the structural flexibility that allows these regulatory mechanisms to operate (Hollenhorst *et al.*, 2011).

## ETS Factors in Cancer

Given the vital cellular processes regulated by ETS factors, it is not surprising that they have also been identified as significant contributors to tumourigenesis. In fact, the founding member of the ETS family (*v-ets*) was discovered as part of the fusion oncogene in the E26 avian acute leukaemia retrovirus, with the novel sequence designated “E26 transformation-specific sequence”. The cellular homologue of *v-ets* was subsequently identified in the chicken (*c-Ets1*), defining the ETS transcription factor family that would grow to 28 members in humans. Comparison of the viral and avian genes revealed that *v-ets* contained unique 5' and 3' RNA sequences, as well as several mutations leading to amino acid substitutions (reviewed in Blair and Athanasiou, 2000).

These early studies of *v-ets* demonstrate several molecular mechanisms (increased expression, mutations and fusions) by which ETS factors can contribute to cancer. Alterations in ETS factor expression or activity drive gene expression programs that

control many of the hallmark features of cancer, including sustained proliferation, altered differentiation, resistance to cell death, invasion, and angiogenesis (Hanahan and Weinberg, 2011). There are various molecular mechanisms underlying ETS factor dysregulation in cancer, including chromosomal translocations, increased or decreased mRNA expression, amplification, mutations in ETS genes or their genomic binding sites, alterations in protein-protein interactions, and changes in protein stability or subcellular localisation (Kar and Gutierrez-Hartmann, 2013; Sizemore *et al.*, 2017). These molecular mechanisms are summarised in Figure 1.17 and examples are provided in Table 1.6. The prevalence and diversity of these alterations in various cancer types underscores the importance of ETS factors in cancer initiation and progression.



**Figure 1.17: Molecular mechanisms of ETS factor activation and inactivation in cancer**

**Figure reproduced (and caption modified) with permission from Nature Publishing Group (Sizemore *et al.*, 2017)**

(A) Chromosomal translocation generates ETS gene fusions in Ewing sarcoma, and prostate, gastric, head and neck, thyroid and breast (secretory) carcinomas. (B) ETS factors are amplified in melanoma (*ETV1*), breast cancer (*ELF3* and *ETV3*), and haematological malignancies. (C) Mutations in the *TERT* promoter generate a novel ETS binding site in up to 70% of melanomas. ETS factor binding to this site results in aberrant *TERT* up-regulation.

(D) Gain-of-function (GOF) p53 mutant proteins interact with ETS2, resulting in enhanced p53 transcriptional activity and altered regulation of target genes. (E) In gastrointestinal stromal tumours, activating mutations in KIT (a receptor tyrosine kinase) increase the phosphorylation and stability of ETV1 in a feed-forward loop. (F) ETV1, ETV4 and ETV5 normally interact with the E3 ubiquitin ligase COP1; however, fusion with TMPRSS2 disrupts the COP1-interaction site, enhancing protein stability approximately 50-fold. Similarly, increased SRC activity in breast cancer leads to phosphorylation of ETS1, inhibiting COP1 binding and proteasomal degradation. ASCL3, achaete-scute homolog 3; CANT1, calcium activated nucleotidase 1; COP1, E3 ubiquitin-protein ligase COP1; DDX1, DEAD-box helicase 1; ERVK6, endogenous retrovirus group K member 6; EWSR1, RNA-binding protein EWS; FUS, RNA-binding protein FUS; HNRPA2B1, heterogeneous nuclear ribonucleoprotein A2/B1; KIT, KIT proto-oncogene receptor tyrosine kinase; NTRK3, neurotrophic receptor tyrosine kinase 3; PDGFR, platelet-derived growth factor receptor; SLC45A3, solute carrier family 45 member 3; TERT, telomerase reverse transcriptase; TMPRSS2, transmembrane protease serine 2; Ub, ubiquitin.

Interestingly, mutations (not involving gene fusions) do not appear to be a major mechanism of dysregulation of ETS factors in solid tumours. However, a recent genomic study has identified ERF mutations in 1-3% of prostate cancers, occurring almost exclusively in tumours that lack the TMPRSS2-ERG gene fusion. The truncating or missense mutations reduce the ERF protein stability and expression levels, leading to an increase in androgen-regulated gene expression and a cellular phenotype closely resembling that driven by TMPRSS2-ERG (Bose *et al.*, 2017). As sequencing technology continues to advance, more ETS factor mutations in solid cancers may be discovered providing further insights into ETS-mediated oncogenesis.

Table 1.6 also illustrates that the oncogenic effects of ETS factors can be highly context-dependent. The same ETS factor can have oncogenic or tumour-suppressive effects depending on the cancer type, which may be related to co-operating oncogenic events or intrinsic features of the cell-of-origin. An example is the contrasting effects of ELF4 over-expression in ovarian and lung cancer cell lines (Seki *et al.*, 2002; Yao *et al.*, 2007). Furthermore, context-dependent effects may extend to cancer sub-types, with alterations in ELF5 expression, for example, having distinct effects in luminal and basal-like breast cancer cells (discussed further below) (Kalyuga *et al.*, 2012).

**Table 1.6: Molecular mechanisms of ETS factor dysregulation in cancer**

<b>Mechanism</b>	<b>Cancer Type</b>	<b>ETS Factors</b>	<b>Additional Information</b>
<b>Chromosomal translocations</b>	Ewing sarcoma	FLI1, ERG, ETV1, ETV4 (fusion with EWSR1)	Fusion of EWS N-terminal with the ETS factor DNA-binding domain
	Prostate cancer	ERG, ETV1, ETV4, ETV5 (fusion with TMPRSS2) ELK4 (with SLC45A3) ETV4 (with DDX1, CANT1)	Approximately 50% of prostate cancers contain a TMPRSS2-ETS fusion, leading to AR-driven expression of the ETS domain-containing fusion protein
	Gastric, head and neck, thyroid, and secretory breast cancer	ETV6 (fusion with NTRK3)	
	Leukaemia	ETV6 (fusions with NTRK3, PDGFRB, JAK2, RUNX1, PAX5, others)	Fusions may involve the ETS and/or PNT domain; PNT domain oligomerisation activates kinase activity in ETS-tyrosine kinase fusions
<b>Increased expression</b>	Breast cancer	ETS1, ETS2, ETV1, ETV4, ETV5	
	Endocrine-resistant luminal breast cancer	ELF5, SPDEF	
	Ovarian cancer	ELF4	
	Prostate cancer	ELK1	
	Leukaemia	ETV7, SPI1, SPIB	
<b>Amplification</b>	Melanoma	ETV1	
	Breast cancer	ETV3 ELF3	ETV3 and ELF3 associated with 1q amplification (Mesquita <i>et al.</i> , 2013)
	Leukaemia and lymphoma	ETS1, ETS2	
<b>Increased transcriptional activity</b>	Breast cancer	ETV1, ETV4, ETV5	ERBB2 overexpression initiates signalling pathways that converge on ETS factors to increase transcriptional activity (as well as expression)
<b>Decreased expression (loss of tumour-suppressive function)</b>	Breast cancer	ELF5, FLI1	
	Urothelial cancer	ELF5	
	Lung cancer	ELF4	
	Prostate cancer	EHF	
	Colon cancer	SPDEF	
	Nasopharyngeal carcinoma	ETV7	
<b>Gene mutations</b>	Leukaemia	SPI1, ETV6	
	Prostate cancer	ERF	Truncating or missense mutations affecting the ETS domain, resulting in reduced protein stability and expression; leads to an increase in AR-driven transcription

<b>Binding site mutations</b>	Melanoma, glioblastoma	Multiple	Mutations in the <i>TERT</i> promoter generate a novel ETS binding site in up to 70% of melanomas (Horn <i>et al.</i> , 2013; Huang <i>et al.</i> , 2013)
<b>Altered protein-protein interactions</b>	Breast, colon, liver, lung, pancreatic, and prostate cancer, osteosarcoma	ETS2, others?	Gain-of-function p53 mutant proteins interact with ETS2; leads to enhanced p53 transcriptional activity and altered regulation of target genes (including many chromatin modifying enzymes), as well as protection of ETS2 from proteasomal degradation (Do <i>et al.</i> , 2012; Zhu <i>et al.</i> , 2015)
	Leukaemia	SPI1	RUNX1 deficiency (due to mutation or translocation events) results in abnormal recruitment of co-repressors to the SPI1 transcriptional complex (Hu <i>et al.</i> , 2011)
<b>Changes in protein stability</b>	Gastrointestinal stromal tumour (GIST)	ETV1	Activating mutations in KIT (a receptor tyrosine kinase) increase the phosphorylation and stability of ETV1
	Prostate cancer	ETV1, ETV4, ETV5 (fusions with TMPRSS2)	Fusion protein does not contain COP1-interacting sites, enhancing stability approximately 50-fold
<b>Changes in subcellular localisation</b>	Breast cancer	ELF3, ELF5	ELF3 cytoplasmic localisation initiates cellular transformation (Prescott <i>et al.</i> , 2004). High cytoplasmic ELF5 associated with poorer prognosis in luminal A tumours (Gallego-Ortega <i>et al.</i> , 2015).

**Table 1.6.** Information sourced from the following reviews unless otherwise indicated: (Findlay *et al.*, 2013; Kar and Gutierrez-Hartmann, 2013; Seth and Watson, 2005; Sizemore *et al.*, 2017). CANT1, calcium activated nucleotidase 1; COP1, E3 ubiquitin-protein ligase COP1; DDX1, DEAD-box helicase 1; ERBB2, erb-b2 receptor tyrosine kinase 2; EWSR1, RNA-binding protein EWS; JAK, Janus kinase 2; NTRK3, neurotrophic receptor tyrosine kinase 3; PAX5, paired box 5; PDGFRB, platelet-derived growth factor receptor beta; RUNX1, runt-related transcription factor 1; SLC45A3, solute carrier family 45 member 3; TERT, telomerase reverse transcriptase; TMPRSS2, transmembrane protease serine 2.



## **The ETS Transcription Factor ELF5**

E74-like factor 5 (ELF5) is an epithelial-specific member of the ETS transcription factor family (Oettgen *et al.*, 1999; Zhou *et al.*, 1998). The ELF5 protein contains both a C-terminal ETS domain (85 amino acids) and an N-terminal PNT domain (83 amino acids). There are 4 ELF5 transcript variants in the NCBI RefSeq database (National Center for Biotechnology Information, 2002), predicted to produce 4 unique proteins. The two full-length transcript variants produce proteins that differ by only 10 N-terminal amino acids (Isoform 1 = 265 amino acids, Isoform 2 = 255 amino acids), while two additional transcripts (Isoforms 3 and 4) are produced by splicing of exons 4 (+/-5) from each of the full-length transcripts; this produces proteins that lack the Pointed domain but retain the ETS domain. The mouse ELF5 Pointed domain has been shown to have strong transactivation activity (Choi and Sinha, 2006), however the mechanisms underlying this activity (for example, protein-protein interactions or post-translational modifications) are unknown.

## **ELF5 Regulates Cell Fate**

A critical function of ELF5 is the regulation of cell fate, beginning with specification of the trophectoderm in the blastocyst (Donnison *et al.*, 2005). Correct spatial and temporal ELF5 expression is also important for normal development of the embryonic lung (Metzger *et al.*, 2008). Prolactin- and progesterone-driven ELF5 expression during pregnancy directs the development of the mammary luminal progenitor cells into oestrogen receptor (ER)- and progesterone receptor (PR)-negative milk-producing cells (Oakes *et al.*, 2008). In normal human tissues, ELF5 is reported to be expressed in the kidney, prostate, lung, mammary gland, salivary gland, placenta and stomach (Lapinskas *et al.*, 2004; Oettgen *et al.*, 1999; Zhou *et al.*, 1998).

In the mammary gland, Elf5 expression gradually increases during pregnancy, peaks during lactation, and falls during involution (Harris *et al.*, 2006). Homozygous deletion of *Elf5* in the mouse mammary gland profoundly inhibits alveolar development and milk production during pregnancy, while ductal development during puberty is unaffected. Conversely, forced Elf5 expression in the mammary epithelial cells results in precocious alveolar development and milk production in virgin mice (Oakes *et al.*, 2008). Elf5 also rescues the failed alveolar development and lactation of the prolactin receptor (PRLR) knockout mouse (Harris *et al.*, 2006). More recently, progesterone was shown to induce Elf5 expression in PR-negative luminal progenitor cells through the paracrine mediator RankL (also known as tumour necrosis factor ligand superfamily 11, Tnfsf11) (Lee *et al.*, 2013). Collectively, these findings have established Elf5 as an

essential downstream effector of the alveolar cell fate program that is initiated by the hormones prolactin and progesterone.

Recent genomic studies have identified additional mechanisms that may contribute to the regulation and activity of Elf5 during mammary development. In a 2016 study by Shin *et al.*, mammary-specific super-enhancers were identified in the lactating mammary gland using ChIP-seq for the transcription factor Stat5a, GR, MED1, and the histone modification H3K27ac. Elf5 motifs were highly enriched within the 440 mammary-specific super-enhancers and Elf5 binding to these sites was subsequently confirmed by ChIP-seq. Further studies focusing on the whey acid protein (*Wap*) super-enhancer showed that it was composed of 3 constituent enhancers, which became progressively activated by Stat5a, Elf5, GR, and nuclear factor 1 (Nfb1) during pregnancy in a hierarchical and synergistic manner. This study suggests that cooperativity between Elf5, Stat5a, and additional transcription factors at enhancer or “super-enhancer” elements contributes to establishment of the secretory lineage. Interestingly, both Stat5 and Elf5 are regulated by enhancer elements that are co-bound by both Stat5 and Elf5, indicating that these two transcription factors co-operate in a positive feedback loop to reinforce the alveolar differentiation program once it is initiated (Metser *et al.*, 2016; Shin *et al.*, 2016).

### **ELF5 in Cancer**

ETS factors are frequently deregulated in cancer through diverse mechanisms. ELF5 was originally described as a tumour suppressor (Zhou *et al.*, 1998), however the role of this protein in cancer is complex and, like other ETS factors, appears to be context-dependent. In prostate cancer, for example, ELF5 has been shown to inhibit TGF- $\beta$ -driven epithelial-mesenchymal transition, by blocking phosphorylation of the TGF- $\beta$  effector protein SMAD3 (Yao *et al.*, 2015). Conversely, ELF5 has been shown to be upregulated in a cell line model of prostate cancer progression involving acquisition of androgen-independence (Xie *et al.*, 2011). Bladder and kidney carcinoma have been associated with loss of ELF5 expression at the protein and RNA level (Lapinskas *et al.*, 2011; Wu *et al.*, 2015), whereas in endometrial carcinoma ELF5 upregulation is associated with higher disease stage (Risinger *et al.*, 2003). ELF5 gene rearrangements have been described in several lung cancer cell lines (Zhou *et al.*, 1998) and a recent case study has described a ZFPM2-ELF5 fusion gene in multicystic mesothelioma (Panagopoulos *et al.*, 2015); however, gene fusions do not appear to be a major mechanism for deregulation of ELF5, in contrast to other ETS factors (Tomlins *et al.*, 2005).

The breast is the most well-studied context for the role of ELF5 in cancer, with microarrays showing increased expression in basal-like and decreased expression in luminal A/B and HER2-enriched breast cancers, suggesting subtype-specific effects (Kalyuga *et al.*, 2012). Analysis of the transcriptional effects of increased Elf5 expression in luminal breast cancer cells (MCF7 and T47D) reveals that it suppresses the luminal oestrogen-responsive phenotype. Conversely, in high-Elf5 basal-like breast cancer cells (HCC1937), sustained Elf5 expression is important in maintaining the basal-like phenotype. In both luminal and basal-like cells, Elf5 suppresses the claudin-low or mesenchymal phenotype. This indicates that Elf5 has both subtype-dependent and subtype-independent effects.

Transient ELF5 expression in luminal cell line models reduces proliferation, invasion, oestrogen receptor-driven transcription, and epithelial-to-mesenchymal transition (Chakrabarti *et al.*, 2012a; Kalyuga *et al.*, 2012). These cell-intrinsic effects of Elf5 contribute to a reduction in metastasis in some *in vivo* models (Chakrabarti *et al.*, 2012a). However, sustained increased ELF5 expression in other contexts is associated with disease progression, for example endocrine-resistant breast cancers, which are dependent on elevated ELF5 for growth in cell line models, and the basal-like subtype of breast cancer (Kalyuga *et al.*, 2012). This suggests that unknown mechanisms also exist whereby ELF5 can promote cancer cell growth and survival. Furthermore, a recent study has demonstrated that increased Elf5 expression in the mouse mammary tumour virus-Polyoma Middle T (MMTV-PyMT) model of luminal breast cancer leads to increased infiltration of myeloid-derived suppressor cells and an increase in lung metastasis (Gallego-Ortega *et al.*, 2015). Therefore, the interaction of breast cancer cells with the immune system and other components of the tumour microenvironment is an additional factor that can contribute to the context-dependent effects of Elf5.

It is hypothesised that the developmental transcriptional programs driven by ELF5 during normal development also function in breast cancer. In this way, an oncogenic event in the luminal progenitor cell (the proposed cell-of-origin for most breast cancers) may be driven towards either a basal-like subtype (by ELF5) or a luminal subtype (by ER). The transcriptional program that predominates may be influenced by how far the luminal progenitor cell has progressed along the differentiation continuum, as well as by the nature of the oncogenic event. Subsequent up-regulation of ELF5 expression in a luminal breast cancer cell may also lead to an oestrogen-insensitive phenotype and the development of endocrine resistance (Gallego-Ortega *et al.*, 2013). Therefore, ELF5 is hypothesised to play an essential role in the transcriptional programs that drive

basal-like breast cancer (which is usually triple-negative) and endocrine-resistant breast cancer, both of which are associated with poor outcome. Furthermore, the indirect effects of ELF5 mediated by recruitment of immune cells enhances the metastatic potential of luminal breast cancer cells, which may further contribute to poor outcomes in a subset of patients with luminal A breast cancer (Gallego-Ortega *et al.*, 2015).

## 1.5 DNA Repair Proteins: A Novel Class of Transcriptional Regulators

One of the proteins identified as an ELF5-interacting protein in this study was DNA-dependent protein kinase catalytic sub-unit (DNA-PKcs) (Chapter 5). DNA-PKcs has well-characterised roles in the repair of double-stranded DNA breaks (DSBs) (Goodwin and Knudsen, 2014). Originally, however, DNA-PKcs was discovered as a transcriptional regulator, phosphorylating the SP1 transcription factor and the CTD of RNA Pol II (Dvir *et al.*, 1992; Jackson *et al.*, 1990). Recent studies have confirmed and expanded the known roles of DNA-PKcs in transcription, which are discussed below to provide a background to Chapter 5.

### DNA Repair and Transcription

There is increasing evidence that the DNA repair and transcriptional machineries are extensively interconnected. On the one hand, DNA damage can impair the transcriptional machinery and lead to errors in transcripts, compromising cell function and survival. On the other hand, the process of transcription itself can cause DNA damage (Fong *et al.*, 2013). A recent study also demonstrated that transcription of non-coding RNA molecules arising at sites of DNA damage may be important for recruitment of the DNA repair machinery (Francia *et al.*, 2012).

The process of transcription is inherently DNA-damaging. As the DNA is threaded through the advancing RNA Pol II complex, the DNA molecule is forced to rotate around its central axis. This results in the generation of torsional stress, promoting the under-twisting of upstream DNA (negative super-coiling) and the over-twisting of downstream DNA (positive super-coiling) (Baranello *et al.*, 2012). The under-twisted upstream strands can separate, exposing single-stranded DNA to chemical modifications and genotoxic insults, and increasing the chance of recombination events (reviewed in Fong *et al.*, 2013). In addition to this non-specific damage, transcription may also specifically induce transient double-stranded DNA breaks that are required for initiation and elongation. The binding of ligand-bound oestrogen receptor (ER) or androgen receptor (AR), for example, results in the recruitment of DNA topoisomerase 2-beta (TOP2B) to regulatory elements (Haffner *et al.*, 2010; Ju *et al.*, 2006). This enzyme functions to relieve torsional stress and DNA tangles by transiently breaking and re-joining the DNA backbone (Deweese and Osherooff, 2009) and is believed to have several functions in transcription. Firstly, the generation of DSBs by TOP2B at sites of nuclear receptor-mediated transcription results in the recruitment of additional

DNA damage-sensing and repair enzymes, including poly(ADP-ribose) polymerase 1 (PARP1) and the DNA-PK complex (which includes DNA-PKcs and two regulatory subunits). The action of these enzymes has been shown to facilitate a permissive chromatin structure for transcriptional initiation, for example through PARP1-initiated histone H1 exchange (Ju *et al.*, 2006). The generation of DSBs may also increase chromatin flexibility, thereby promoting promoter-enhancer communication (Fong *et al.*, 2013). Finally, the generation of DSBs by TOP2B during transcriptional elongation relieves torsional stress on the DNA, thereby facilitating the passage of RNA Pol II and helping to limit transcription-associated DNA damage (Ju and Rosenfeld, 2006).

A large number of classic DNA repair proteins, representing diverse repair pathways, have now been shown to have additional roles in transcription (see Fong *et al.*, 2013 for an excellent review). Examples include thymine DNA glycosylase (TDG), the helicases ERCC excision repair proteins 2 and 3 (ERCC2 and ERCC3), the endonucleases ERCC4 and ERCC5, poly(ADP-ribose) polymerase 1 (PARP1) and, of course, DNA-PKcs. In addition, a number of DNA-binding sequence-specific transcription factors have been shown to be required for DNA repair, for example the nuclear receptor subfamily 4 group A members (NR4A1-4) (Malewicz *et al.*, 2011) and the zinc finger transcription factor leukaemia/lymphoma-related factor (LRF) (Liu *et al.*, 2015). The association between transcription factors and DNA repair proteins may have originally developed to protect genomic DNA from transcription-induced damage. Subsequently, however, this association appears to have evolved, with the diverse enzymatic capabilities of DNA repair proteins (including glycosylases, helicases, nucleases, kinases, and other ATPase activities) also being utilised for numerous regulatory roles in transcription (Fong *et al.*, 2013). Indeed, the enzymatic activities of transcription-associated DNA damage and repair proteins (including TOP2B, PARP1 and DNA-PKcs) have been shown to be essential for gene regulation by a number of transcription factors, particularly those associated with developmental or stimulus-induced gene regulation (Brenner *et al.*, 2011; Foulds *et al.*, 2013; Goodwin *et al.*, 2015; Haffner *et al.*, 2010; Ju *et al.*, 2006; Medunjanin *et al.*, 2010b). The mechanisms by which these proteins regulate transcription are varied and include modulation of transcription factor activity (for example, through post-translational modifications), regulation of co-factor dynamics, and alterations in epigenetic modifications and chromatin structure. The transcriptional roles of these proteins are believed to be distinct from their DNA repair roles, as demonstrated by the lack of transcriptional effects when essential downstream DNA repair factors are depleted (for example, DNA ligase 4 or its partner XRCC4, which are required for the final DNA ligation step in the

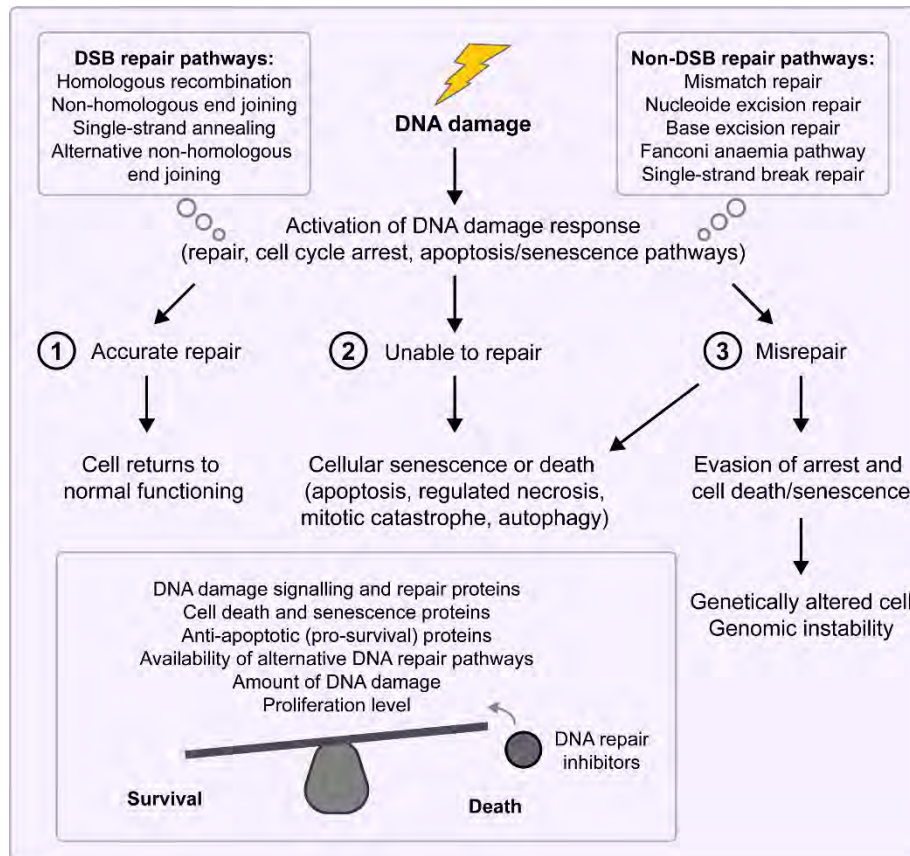
non-homologous end joining repair pathway) (Brenner *et al.*, 2011).

### **The DNA-Damage Response**

DNA repair proteins canonically function in the complex signalling cascade that is activated by DNA damage known as the DNA-damage response (DDR) (Figure 1.18). Cells are constantly exposed to endogenous and exogenous agents that damage DNA and threaten genomic integrity. It is estimated, for example, that a single cell may experience up to 100,000 spontaneous DNA lesions per day (Hoeijmakers 2009). Amazingly, the vast majority of these lesions are effectively repaired through the use of various pathways, which specialise in specific types of DNA damage. Double-stranded DNA breaks (DSBs) are one of the most harmful forms of DNA damage and there are two main pathways responsible for their repair in human cells - non-homologous end joining (NHEJ) and homologous recombination (HR). As the name suggests, non-homologous end joining directly ligates broken DNA ends without a requirement for homology; it is rapid but somewhat error-prone, occurs throughout the cell cycle and is the primary method of DSB repair in humans. Homologous recombination, in contrast, uses the sister chromatid as a template for repair during the S and G2 phases of the cell cycle and is an extremely accurate method of DSB repair. Other types of DNA damage and their associated repair pathways include mispaired DNA bases (mismatch repair pathway), chemical modifications to DNA bases (base excision repair), pyrimidine dimers and intrastrand crosslinks (nucleotide excision repair), and single-strand DNA breaks (single-strand break repair) (Ciccia and Elledge, 2010; Lord and Ashworth, 2012).

The DDR is initiated by molecular sensors of DNA damage, such as PARP1 and the MRE11/RAD50/NBS1 (MRN) complex (Lavin, 2007). These sensor proteins recruit and activate the PI3K-related kinases ataxia telangiectasia mutated serine/threonine kinase (ATM) and ataxia telangiectasia and RAD3-related serine/threonine kinase (ATR), which are the primary signalling kinases involved in the DDR. The DNA-PK complex is also recruited to sites of DSBs by its regulatory subunits, where it is primarily involved in the recruitment and activation of a smaller group of NHEJ repair proteins (Ciccia and Elledge, 2010). The DDR triggers a cascade of cellular responses, involving rapid phosphorylations and slower transcriptional events, ultimately regulating the interconnected pathways of DNA repair, cell cycle arrest and cell death. Cell cycle arrest is an important mechanism that allows time for DNA repair and prevents the propagation of damaged DNA. ATM, for example, phosphorylates multiple targets including CHK1, CHK2 and p53, resulting in a G1/S or G2 arrest. The DDR pathway

can also activate pro-death proteins, such as Fas ligand (FASL), Fas receptor (FASR), BIM, BAX, PUMA and NOXA. Therefore, the DDR activates both pro-survival (repair) and pro-death pathways. An important determinant of cell fate in response to DNA damage is p53, which regulates proteins involved in both pathways (Roos *et al.*, 2016).



**Figure 1.18: The DNA-damage response**

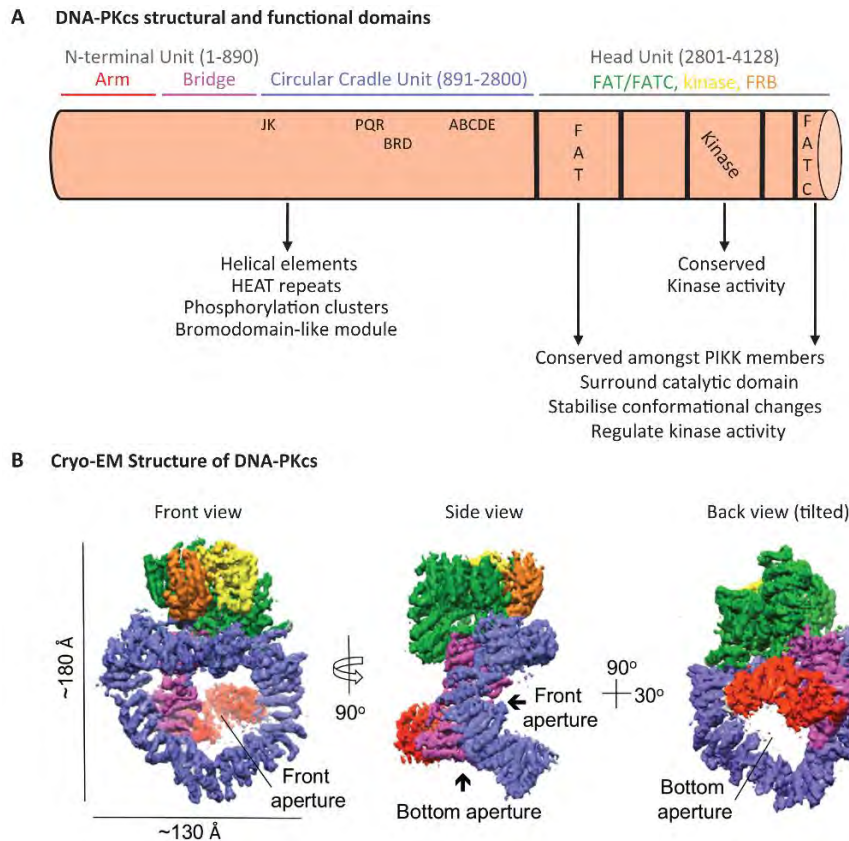
In order to protect genomic integrity, multiple pathways exist for recognising and repairing various types of DNA damage. DNA damage is detected by molecular sensor proteins, which recruit and activate the PI3K-related kinases ATM and ATR to initiate the DNA-damage response (DDR). Rapid phosphorylation events transmit and amplify the damage signal, culminating in the regulation of DNA repair, cell cycle arrest and cell death/senescence pathways. There are three main outcomes of DNA damage: 1. Full and accurate repair of damage, 2. Cell death or senescence due to an inability to repair damage, and 3. Misrepair, leading to cell death or (in the case of mutation of essential regulators of cell death such as p53) survival of a genomically altered cell. A number of factors influence the balance between survival and death following DNA damage, including the expression and activity of DDR and repair proteins, the amount of DNA damage, and the proliferation level. However, the exact mechanisms by which the cell quantifies DNA damage and decides between these different fates are still not well understood (Roos *et al.*, 2016).



In general, there are three possible outcomes arising from initiation of the DDR (labelled 1-3 in Figure 1.18), with cell survival and cell death poised in a delicate balance. Firstly, the activation of repair pathways may allow the cell to fully and accurately repair the damage and to return to normal functioning. Secondly, the cell may be unable to repair the damage (for example, if the repair fails or if the amount of damage exceeds the repair capacity of the cell). In this case, cell senescence or death pathways are definitively activated, preventing the propagation of damaged DNA. Cell death in response to DNA damage may occur through multiple mechanisms, including apoptosis, regulated necrosis, mitotic catastrophe or autophagy (Surova and Zhivotovsky, 2013). Thirdly, damage may be misrepaired, leading to an alteration in genomic DNA, which may also trigger cell death pathways. However, the acquisition of mutations that affect cell cycle and cell death regulators (for example, p53) can allow evasion of cell death and continued proliferation, resulting in a population of cells with potentially oncogenic genomic alterations (Khanna, 2015).

### **The DNA-Dependent Protein Kinase Catalytic Sub-Unit**

The DNA repair protein DNA-PKcs was identified in this project as an ELF5-interacting protein. DNA-PKcs is a large (469 kDa) serine/threonine protein kinase encoded by the *PRKDC* gene on human chromosome 8. As a member of the phosphatidylinositol 3-kinase-related kinase (PIKK) family, it is related to the DNA-damage kinases ataxia telangiectasia mutated (ATM) and ataxia telangiectasia and Rad3 related (ATR), as well as the mammalian target of rapamycin (mTOR) (Goodwin and Knudsen, 2014). The structural and functional domains of DNA-PKcs are shown in Figure 1.19. The most recent structures of the DNA-PKcs protein shows that it forms three main units - the N-terminal region (with an arm and a bridge), the circular cradle and the C-terminal head, which includes the kinase domain (Figure 1.19B) (Sharif *et al.*, 2017; Sibanda *et al.*, 2017). The DNA-PKcs structure contains two openings (front and bottom apertures) and it is proposed that double-stranded DNA passes through these gaps (Sharif *et al.*, 2017). Phosphorylation of DNA-PKcs (by DNA-PKcs itself or by other kinases such as ATM, ATR and AKT1) is an important regulatory mechanism and causes significant conformational changes (reviewed in Jette and Lees-Miller, 2015). Current evidence suggests that the N-terminal is responsible for the interaction with DNA and communicates allosteric information on DNA-binding to the kinase domain in the head unit (Sibanda *et al.*, 2017). Ku80 also interacts with DNA-PKcs in the N-terminal arm region (Sharif *et al.*, 2017). Interestingly, DNA-PKcs is the only known kinase that is reliant on DNA binding for activity (Neal and Meek, 2011).



**Figure 1.19: Functional domains and structure of DNA-PKcs**

**Figures adapted with permission from (A) American Association for Cancer Research (Goodwin and Knudsen, 2014) and (B) Proceedings of the National Academy of Sciences of the United States of America (Sharif et al., 2017)**

(A) DNA-PKcs is composed of multiple structural and functional domains. The large N-terminal and circular cradle units contain helical elements, HEAT repeats and several regulatory phosphorylation clusters (JK, PQR and ABCDE). A novel bromodomain-like module (BRD) has also been recently identified (amino acids 2070-2200) (Wang *et al.*, 2015b). Phosphorylation of key residues (for example, serine 2056 in the PQR cluster and threonine 2609 in the ABCDE cluster) regulate DNA-PKcs complex formation and activity. N-terminal conformation changes due to DNA binding also regulate DNA-PKcs function. The FAT and FATC domains are conserved amongst the phosphatidylinositol-3-kinase-related kinase (PIKK) family and function to stabilise conformational changes in the catalytic kinase domain and regulate enzymatic activity. The kinase domain also contains an FRB domain.

(B) Structural units of DNA-PKcs, with colour-coding corresponding to labels in A. The N-terminal unit consists of the arm (red) and bridge (pink). The large circular cradle is shown in purple. The FAT and FATC (green), kinase (yellow) and FRB (orange) domains comprise the head unit. The structure contains two openings (front and bottom apertures) and DNA is proposed to pass through these gaps. FAT, domain with homology to FAT, ATM and transformation/transcription domain-associated (TRRAP) proteins; FATC, FAT at the extreme C-terminus domain; FRB, FKBP12-rapamycin-binding; HEAT repeat, Huntingtin, Elongation Factor 3, PP2A and TOR1 repeat; Nt, N-terminus; Ct, C-terminus.

## The Role of DNA-PKcs in DNA Repair

The most well-characterised role of DNA-PKcs is in the repair of double-stranded DNA breaks (DSBs) through non-homologous end joining (NHEJ). It is also involved in the regulation of homologous recombination (HR) and influences the selection of DSB repair pathway (Neal and Meek, 2011; Zhou *et al.*, 2017). DSBs may arise due to physiological processes (for example, due to the generation of reactive oxygen species or as a result of a stalled DNA replication fork) or exogenous mechanisms (for example, exposure to ionising radiation or chemotherapy drugs such as doxorubicin). NHEJ is the primary repair pathway for DSBs in humans and is initiated by the high-affinity binding of x-ray repair cross-complementing proteins 5 and 6 (XRCC5 and 6, also known as Ku70 and Ku80 or 86) to the ends of broken DNA. The interaction of DNA-PKcs with both DNA and the DNA-bound Ku heterodimer increases its kinase activity (5-10-fold) through a largely unknown mechanism and results in regulatory phosphorylation of DNA-PKcs itself, the histone variant H2AX, ATM, and various DNA repair proteins (reviewed in Goodwin and Knudsen, 2014; Jette and Lees-Miller, 2015; Zhou *et al.*, 2017). In addition, DNA-PKcs binding brings the broken DNA ends into close proximity and protects them from degradation by nucleases (DeFazio *et al.*, 2002; Weterings *et al.*, 2003). It also regulates the arrest and eviction of elongating RNA Pol II upon encountering a DSB to prevent the production of error-prone transcripts (Pankotai *et al.*, 2012). The Ku heterodimer interaction is essential for all known DNA repair functions of DNA-PKcs and together these proteins form the DNA-PK complex. Various factors are recruited to process the DNA ends, which are then-joined by DNA ligase IV (Mahaney *et al.*, 2009).

Defects in DNA-PKcs expression or activity result in inefficient repair of double-stranded DNA breaks. In early studies, rodent cell lines with inactivating mutations in DNA-PKcs or Ku80, for example, were shown to be highly sensitive to death caused by ionising radiation (reviewed in Smith and Jackson, 1999). Cultured primary mouse embryonic fibroblasts with DNA-PKcs deficiency exhibit an increased number of chromosomal abnormalities, indicating that DNA-PKcs is important in the repair of endogenous, as well as exogenous, DNA damage (Ferguson *et al.*, 2000). The kinase activity of DNA-PKcs is required for its role in DNA repair and pharmacological inhibition (for example, using the non-specific PI3-kinase inhibitor wortmannin or more specific inhibitors such as NU7441) causes similar defects in DSB repair (Kurimasa *et al.*, 1999; Zhao *et al.*, 2006a).

*In vivo*, loss of DNA-PKcs expression or activity also results in defective DNA repair. The severe combined immunodeficiency (SCID) mouse has a truncating mutation in DNA-PKcs, causing increased sensitivity to radiation and a reduced ability to repair double-stranded DNA breaks. The lack of DNA-PKcs is also responsible for the canonical immunodeficiency in the SCID mouse, as the NHEJ pathway is essential for V(D)J recombination in developing T and B lymphocytes (reviewed in Collis *et al.*, 2005). In humans, the first homozygous germline mutation in DNA-PKcs was recently identified and was shown to be a novel cause of human severe combined immunodeficiency as well as a radiosensitive cellular phenotype (van der Burg *et al.*, 2009). Interestingly, this mutation affected DNA-PKcs-mediated activation of the nuclease Artemis and did not alter the kinase activity of DNA-PKcs. The same mutation was subsequently identified in two additional patients with an immunodeficiency, granuloma and autoimmunity syndrome. The autoimmune manifestations were attributed to a likely impairment of AIRE-mediated transcription of tissue-specific antigens in the thymus (Mathieu *et al.*, 2015). The final known case of germline DNA-PKcs mutation was described in a patient with SCID and severe neurological defects, in which compound heterozygous mutations resulted in substantially decreased but still detectable protein expression and activity (Woodbine *et al.*, 2013). Intriguingly, no cases of complete loss of DNA-PKcs expression or activity have been described in humans, despite occurring spontaneously in mice, horses and dogs with SCID. In addition, DNA-PKcs mutations in other species are not associated with neurological abnormalities. These inter-species differences suggest that human cells may have a unique dependence on DNA-PKcs, which may be related to DNA repair as well as functions outside this canonical role.

### **Emerging Roles for DNA-PKcs**

It is becoming increasingly recognised that DNA-PKcs has a number of important functions outside DNA repair. These include functions in mitosis, telomere maintenance, ageing, metabolism, immunity, the hypoxic response and hormone signalling (Table 1.7). An emerging theme from these roles is that DNA-PKcs can function in multiple capacities as a transcriptional regulator.

**Table 1.7: Emerging roles for DNA-PKcs**

Process	DNA-PKcs Roles	References
Mitosis	Localises to the spindle apparatus and is required for mitotic entry, normal chromosome segregation and cell cycle progression; depletion or inhibition results in chromosome misalignments and delays in mitotic progression. Mitotic phosphorylation events occur independently of DNA damage and do not require the Ku70/80 heterodimer, suggesting a unique mechanism of DNA-PKcs activation.	(Douglas <i>et al.</i> , 2014; Lee <i>et al.</i> , 2011b)
	Functions in the ATR signalling pathway to phosphorylate replication protein A (RPA) and inhibit entry into mitosis during replication stress (the slowing or stalling of DNA replication forks commonly driven by oncogenes)	(Liu <i>et al.</i> , 2012)
Telomere maintenance	Protects telomere ends (particularly newly synthesised telomeres in mitosis), preventing end-to-end chromosomal fusions as well as fusions of telomeres with double-stranded DNA breaks	(Bailey <i>et al.</i> , 2004; Bailey <i>et al.</i> , 1999; Gilley <i>et al.</i> , 2001; Zhang <i>et al.</i> , 2016a)
	Co-operates with telomerase to maintain telomere length, suggesting a possible role for DNA-PKcs deficiency in ageing	(Espejel <i>et al.</i> , 2002; Espejel <i>et al.</i> , 2004)
Metabolism	Increased DNA-PKcs activity in skeletal muscle with advancing age; culminates in decreased AMP-activated protein kinase (AMPK) activity, which has been linked to age-associated obesity, insulin resistance, mitochondrial loss and a reduction in physical fitness	(Park <i>et al.</i> , 2017)
	Phosphorylates and activates AMPK during glucose deprivation	(Amatya <i>et al.</i> , 2012)
	Phosphorylates and activates the transcription factor upstream stimulatory factor 1 (USF1) in response to feeding, leading to increased transcription of hepatic lipogenic genes such as fatty acid synthase	(Wong <i>et al.</i> , 2009)
Innate immunity	Acts as a sensor of foreign viral DNA in the cell cytoplasm, activating cytokine transcription through interferon regulatory factor 3 (IRF3); depletion of DNA-PK reduces the transcriptional response to DNA (but not RNA) viruses. Does not require the kinase activity of DNA-PKcs, although DNA-PKcs has been shown to phosphorylate and stabilise IRF3. Viral defence mechanisms target DNA-PK components for degradation and inhibition.	(Ferguson <i>et al.</i> , 2012; Karpova <i>et al.</i> , 2002; Parkinson <i>et al.</i> , 1999; Peters <i>et al.</i> , 2013)
Adaptive immunity	Required for V(D)J recombination to generate B- and T-cell receptors	(Collis <i>et al.</i> , 2005)
	Targets the transcription factor autoimmune regulator (AIRE) to chromatin by tethering AIRE to DNA; facilitates the expression of normally silent tissue-specific antigens in thymic cells to allow the development of immunological self-tolerance	(Zumer <i>et al.</i> , 2012)
Hypoxic response	Activated by hypoxia and stabilises the transcription factor hypoxia inducible factor 1 alpha subunit (HIF1A)	(Bouquet <i>et al.</i> , 2011; Um <i>et al.</i> , 2004)
Hormone signalling	Interacts with and regulates multiple nuclear hormone receptors, including ER, AR, PR, GR and TR (more detailed information in Table 1.9)	See Table 1.8

Some of the earliest studies on DNA-PKcs demonstrated that it phosphorylates the DNA-bound transcription factor SP1 (Jackson *et al.*, 1990) as well as the CTD of RNA Pol II and other members of the pre-initiation complex (Chibazakura *et al.*, 1997; Dvir *et al.*, 1992). Recently, interest in the transcriptional roles of DNA-PKcs has been renewed, fuelled by the understanding that DNA repair proteins and transcription factors are intimately connected. DNA-PKcs has since been shown to interact with and regulate the activity of multiple transcription factors from diverse families (Table 1.8), as well as additional members of the basal transcriptional machinery (Bunch *et al.*, 2014). Through these interactions, DNA-PKcs is involved in multiple stages of transcriptional regulation, including initiation, pause release and elongation. Most relevant to this project is the discovery that DNA-PKcs can regulate the steroid receptors oestrogen receptor (ER) and androgen receptor (AR), as well as several ETS transcription factors including ETS1 and prostate cancer-associated ETS fusion proteins. Furthermore, a number of transcription factors can reciprocally regulate DNA-PKcs expression and activity.

**Table 1.8: Examples of transcription factors interacting with DNA-PKcs**

Transcription factor	Functional relationship between DNA-PKcs and transcription factor	Additional interactors	References
<b>ETS transcription factors</b>			
ETS proto-oncogene 1 transcription factor (ETS1)	Phosphorylates ETS1 <i>in vitro</i> No effect on transcriptional activity (reporter assay)	Ku70, Ku80	(Choul-li <i>et al.</i> , 2009)
ETS fusion proteins	Interaction requires ERG DBD Complex recruited to DNA by ERG fusion protein and is required for TMPRSS2-ERG transcriptional activity (reporter assay and gene expression) Inhibits NHEJ activity of DNA-PKcs	Ku70, Ku80, PARP1	(Brenner <i>et al.</i> , 2011; Chatterjee <i>et al.</i> , 2015)
Various ETS factors (ERG, ETS1, ETV1, SPI1)	Required for ERG and ETV1 transcriptional activity (gene expression)	Ku70, Ku80, PARP1	(Brenner <i>et al.</i> , 2011)
<b>Nuclear receptors</b>			
Androgen receptor (AR)	Increases transcriptional activity dependent on kinase activity (gene expression) Activated AR positively regulates DNA-PKcs expression and activation Interaction does not require presence of DNA or the AR LBD	Possibly Ku70, Ku80 (Mayeur <i>et al.</i> , 2005)	(Goodwin <i>et al.</i> , 2015; Goodwin <i>et al.</i> , 2013; Mayeur <i>et al.</i> , 2005)
Estrogen receptor (ERS1)	Phosphorylates ER at serine 118 Phosphorylates and activates ER co-activators (SRC3, MED1) Phosphorylates and dismisses transient ER co-repressors from DNA (NRIP1) Increases transcriptional activity and transcriptional response to E2 stimulation (reporter assay and gene expression) Inhibits ER ubiquitination, resulting in increased ER protein stability Activated ER positively regulates DNA-PKcs	Ku70, Ku80, PARP1, TOP2B	(Foulds <i>et al.</i> , 2013; Ju <i>et al.</i> , 2006; Medunjanin <i>et al.</i> , 2010a; Medunjanin <i>et al.</i> , 2010b)

	expression and activation Interaction requires ER B-domain (AF1)		
Glucocorticoid receptor (rodent) (GR)	Phosphorylates rodent GR (Ku and DNA-dependent)	Ku70, Ku80	(Giffin <i>et al.</i> , 1997)
Nuclear receptor subfamily 4 group A 1, 2, 3 (NR4A1, 2, 3)	Phosphorylates NR4A receptors No effects on transcriptional activity Required for NR4A receptors to function in DNA repair	PARP1	(Malewicz <i>et al.</i> , 2011)
Progesterone receptor (PR)	Phosphorylates PR (chicken and human) Increases transcriptional activity (reporter assay using chimaeric PR/ER protein and gene expression) Interaction requires PR DBD	Ku70, Ku80, PARP1	(Sartorius <i>et al.</i> , 2000; Trevino <i>et al.</i> , 2016; Weigel <i>et al.</i> , 1992)
Thyroid hormone and retinoid X receptor heterodimer, unliganded (TR-RXR)	Phosphorylates HDAC3 (part of NCOR1/2 co-repressor complex) and increases its activity, enhancing unliganded TR-RXR transcriptional repression (reporter assay) Interaction is inhibited by binding of thyroid hormone to TR	Ku70, Ku80, PARP1	(Jeyakumar <i>et al.</i> , 2007)
<b>Forkhead transcription factors</b>			
Forkhead box A1 (FOXA1)	DNA repair complex in combination with FOXA1 contributes to targeted DNA demethylation (note: DNA-PKcs interaction was identified using one method but could not be validated by subsequent methods)	Ku70, Ku80, PARP1, POLB, LIG3, XRCC1	(Zhang <i>et al.</i> , 2016c)
Forkhead box A2 (FOXA2)	Phosphorylates FOXA2 (serine 283 of rodent FOXA2 C-terminal to DBD) Mutation of S283 decreases FOXA2 transcriptional activity (reporter assay) Interaction requires DBD of FOXA2	Ku70	(Nock <i>et al.</i> , 2009)
<b>Basic helix-loop-helix transcription factors</b>			
MYC proto-oncogene bHLH transcription factor (MYC)	Phosphorylation of MYC <i>in vitro</i> Stabilisation of MYC protein, although this may be an indirect effect of DNA-PKcs expression Conflicting reports on effects of MYC on DNA-PKcs-mediated DNA repair Reciprocal activation of gene expression	Ku70, Ku80	(An <i>et al.</i> , 2005; An <i>et al.</i> , 2008; Cui <i>et al.</i> , 2015; Iijima <i>et al.</i> , 1992; Koch <i>et al.</i> , 2007; Li <i>et al.</i> , 2012; Zhou <i>et al.</i> , 2014b)
Upstream stimulatory factor 1 (USF1)	Phosphorylates USF1, increasing its transcriptional activity (gene expression)	Ku70, Ku80, PARP1, TOP2B, PP1	(Wong <i>et al.</i> , 2009)
<b>Zinc finger (C2H2) transcription factors</b>			
Leukemia/ lymphoma-related factor (LRF)	LRF stabilises DNA-PKcs at sites of DNA damage and activates its kinase activity	Ku70, Ku80	(Liu <i>et al.</i> , 2015)
Snail family transcriptional repressor 1 (SNAI1)	Phosphorylates SNAI1, increasing protein stability and transcriptional activity (reporter assay) Phosphorylated SNAI1 may inhibit DNA-PKcs kinase activity	Not examined	(Kang <i>et al.</i> , 2013; Pyun <i>et al.</i> , 2013)
Snail family transcriptional repressor 2 (SNAI2 or SLUG)	Not examined	Not examined	(Kang <i>et al.</i> , 2013)
Sp1 transcription factor (SP1)	Phosphorylates SP1 Phosphorylation requires DNA and SP1 binding to DNA	Not examined	(Jackson <i>et al.</i> , 1990)

Other transcription factor families (SAND, HSF, bHLH-PAS, IRF, bZIP, HMG, POU)			
Autoimmune regulator (AIRE)	Phosphorylates AIRE and activates transcriptional activity ( <i>in vitro</i> ) Targets AIRE to sites of transcription; not dependent on kinase activity of DNA-PKcs Required for AIRE transcriptional activity <i>in vivo</i> (gene expression and reporter assay)	Ku70, Ku80, PARP1, TOP2A, H2AX, SUTP16H (FACT140)	(Abramson <i>et al.</i> , 2010; Liiv <i>et al.</i> , 2008; Zumer <i>et al.</i> , 2012)
Heat shock factor 1 (HSF1)	Phosphorylates HSF1 HSF1 stimulates DNA-PKcs activity ( <i>in vitro</i> ) Interaction does not require the presence of DNA but is enhanced by doxorubicin treatment	Ku70, Ku80	(Evert <i>et al.</i> , 2013; Huang <i>et al.</i> , 1997)
Hypoxia inducible factor 1 alpha subunit (HIF1A)	DNA-PKcs is activated by hypoxia Stabilises HIF1A protein Stabilisation dependent on kinase activity and presence of Ku70/80 sub-units Unclear if these effects are mediated by a direct protein-protein interaction	Not examined	(Bouquet <i>et al.</i> , 2011; Um <i>et al.</i> , 2004)
Interferon regulatory factor 3 (IRF3)	Phosphorylates IRF3, increasing nuclear retention and protein stability	Not examined	(Ferguson <i>et al.</i> , 2012; Karpova <i>et al.</i> , 2002)
Jun proto-oncogene, AP-1 transcription factor subunit (JUN)	Phosphorylates JUN ( <i>in vitro</i> ) Phosphorylation enhanced by binding of JUN to cognate DNA sequences	Not examined	(Bannister <i>et al.</i> , 1993)
Lymphoid enhancer binding factor 1 (LEF1)	Not examined	PARP1	(Shimomura <i>et al.</i> , 2013)
POU domain, class 2, transcription factor 1 (POU2F1 or OCT1)	Phosphorylates POU2F1 Increases protein stability but decreases transcriptional activity Effects dependent on POU2F1 interaction with Ku sub-units	Ku70, Ku80	(Schild-Poulter <i>et al.</i> , 2001; Schild-Poulter <i>et al.</i> , 2007; Schild-Poulter <i>et al.</i> , 2003)

**Table 1.8.** AF1, activation function 1 (ligand-independent transactivation domain); bHLH-PAS, basic helix-loop-helix- Per/ARNT/Sim; bZIP, basic leucine zipper; DBD, DNA-binding domain; E2, oestradiol; ERG, transcriptional regulator ERG; ETV1, ETS translocation variant 1; H2AX, H2A histone family member X; HMG, high mobility group; HSF, heat shock factor; IRF, interferon regulatory factor; LBD, ligand-binding domain; LIG3, DNA ligase 3; MED1, mediator complex subunit 1; NCOR1/2, nuclear co-repressor 1 or 2; NHEJ, non-homologous end joining; NRIP1, nuclear receptor interacting protein 1; PARP1, poly(ADP-ribose) polymerase 1; POLB, DNA polymerase beta; POU, Pit1, Oct1/2, Unc86 (protein domain); PP1, protein phosphatase 1; SAND, Sp100, AIRE1, NucP41/75, DEAF1 (protein domain); SPI1, Spi1 proto-oncogene; SRC3, steroid receptor co-activator 3; SUPT16H, SPT16 homolog, facilitates chromatin remodelling subunit (also known as FACT140); TMPRSS2-ERG, transmembrane protease serine 2 fusion with ETS transcription factor ERG, commonly occurring in prostate cancer; TOP2A, DNA topoisomerase 2-alpha; TOP2B, DNA topoisomerase 2-beta; XRCC1, x-ray repair cross-complementing protein 1.



## DNA-Damage Response Proteins in Cancer

DNA-damage response proteins play multiple, and sometimes paradoxical, roles in the development and progression of cancer. The DDR is essential to the maintenance of genomic stability and is almost always dysregulated in cancer, presenting both therapeutic challenges and opportunities (Figure 1.20). The recent recognition that these proteins also frequently function as transcriptional regulators adds further complexity to this issue.

In normal cells, the DNA-damage response functions to prevent genomic instability (defined as the failure to transmit DNA accurately). Genomic instability is a characteristic of essentially all cancers and two main mechanisms are known to contribute to its development. Firstly, genomic instability may arise due to mutations in DNA-damage signalling or repair proteins that impair their expression or function, particularly in hereditary cancers. Early studies of cancer predisposition syndromes (for example, xeroderma pigmentosum and Lynch syndrome) contributed to the understanding of defective DNA repair in cancer (reviewed in Jeggo *et al.*, 2016). There are now many examples of hereditary syndromes with an increased risk of cancer known to be caused by mutations in DNA-damage signalling and repair proteins (see Ciccia and Elledge, 2010 for an excellent review). One example is familial breast cancer, which can be caused by heterozygous mutations in a number of DDR proteins involved in homologous recombination, including breast cancer susceptibility type 1 and type 2 proteins (BRCA1, BRCA2), partner and localiser of BRCA2 (PALB2), checkpoint kinase 2 (CHK2) and ATM (reviewed in Ciccia and Elledge, 2010). Mutations or epigenetic silencing of these genes can also arise in sporadic cancers; in high-grade serous ovarian carcinoma, for example, 3% of cases have somatic *BRCA1* mutations and 12% have hypermethylation and silencing of the *BRCA1* promoter (Cancer Genome Atlas Research Network, 2011).

The second mechanism of genomic instability in cancer is oncogene-driven replication stress. The phenomenon of oncogene-driven genomic instability was first identified in murine fibroblasts overexpressing the HRas GTPase proto-oncogene (HRAS) (Denko *et al.*, 1994). Subsequent studies have proposed that oncogene-driven replication stress (the deregulation of DNA replication, resulting in the slowing or stalling of DNA replication forks) underpins this increase in genomic instability (Halazonetis *et al.*, 2008). Recent next-generation sequencing studies suggest that oncogene-driven replication stress, rather than early mutations in DDR proteins, may be the more prominent mechanism of genomic instability in sporadic cancers (Negrini *et al.*, 2010).

The genomic instability driven by oncogenes activates the DDR, which acts as a natural barrier to cancer progression by inducing cell death or senescence. However, the loss of DNA-damage response proteins such as p53 (possibly due to genetic alterations arising from replication stress) allows evasion of these DDR-mediated barriers and is an essential step in the progression to uncontrolled proliferation and cancer (Bartkova *et al.*, 2005; Gorgoulis *et al.*, 2005). Therefore, oncogene-induced replication stress can account for the development of genomic instability in human cancers as well as the high incidence of p53 mutations, which are selected for due to the ability of these cells to bypass the DDR safeguards (Halazonetis *et al.*, 2008).

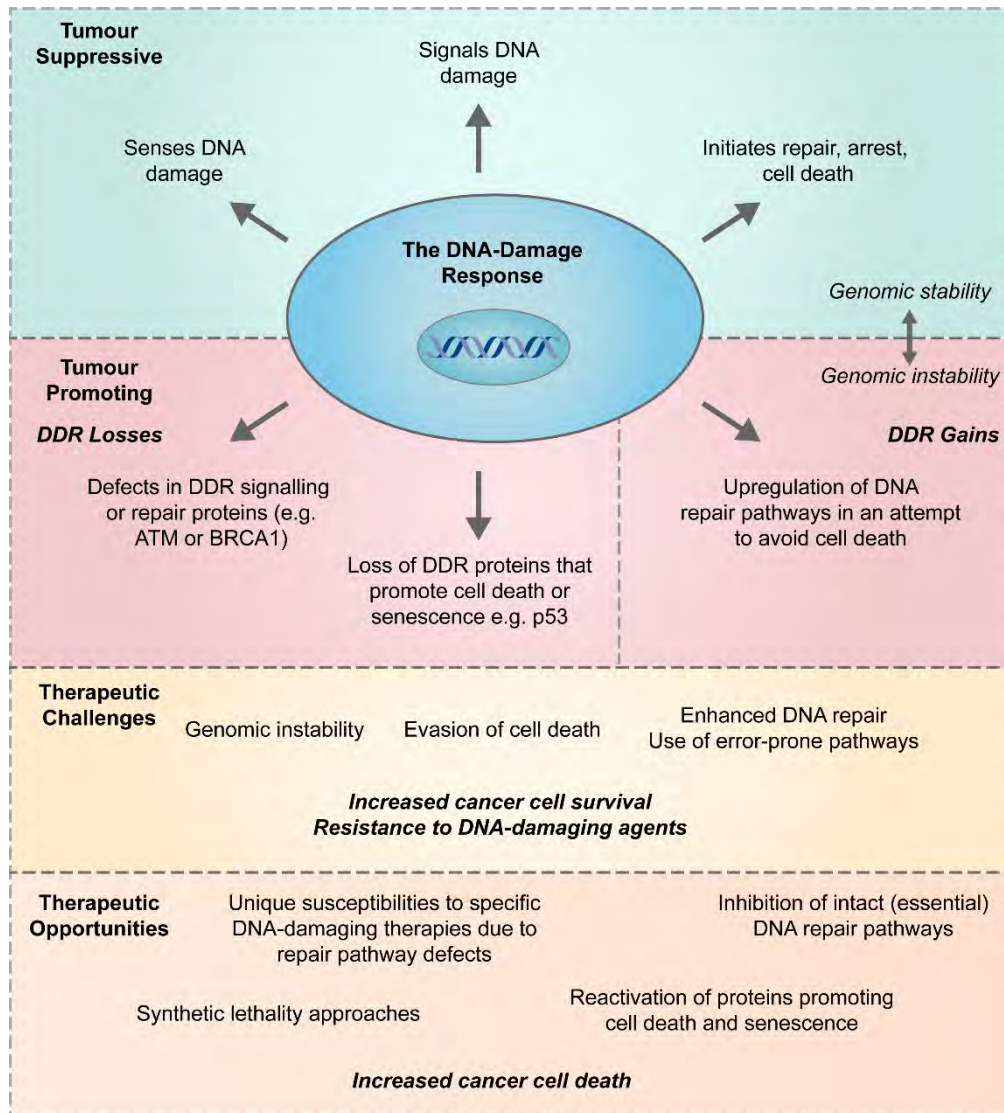
The above discussion indicates that DNA-damage signalling and repair proteins are tumour-suppressive in normal cells and early cancerous lesions. However, somewhat counterintuitively, functional DDR proteins such as checkpoint kinase 1 (CHK1) and DNA-PKcs have been shown to be frequently upregulated in cancer cells and essential for continued survival (reviewed in Khanna, 2015). This is because these proteins promote cell cycle arrest and DNA repair, allowing the cancer cell to avoid lethal pathways such as mitotic catastrophe (broadly defined by Vitale *et al.*, 2011 as the induction of cell senescence or death (by apoptosis or necrosis) resulting from a failure of mitosis ). Therefore, DDR proteins can function either as tumour suppressors or tumour promoters, depending on the context (for example, the presence of other genomic alterations and tissue type). Furthermore, cancer therapies such as radiotherapy and chemotherapy frequently aim to induce an even higher level of DNA damage to overwhelm the repair capacity of the cell and induce cell death (Khanna, 2015). Therefore, high levels of DNA repair proteins such as CHK1 and DNA-PKcs can promote resistance to cancer therapies.

Dysregulation of the DDR pathway in cancer presents therapeutic challenges but also therapeutic opportunities. Firstly, there is the challenge of treatment resistance, with high levels of DNA repair proteins making the DNA-damage-induced shift from cell survival to cell death more challenging. Secondly, ongoing genomic instability in the cancer cell promotes adaptability through additional genomic alterations that can compensate for previous genomic changes or the effects of treatments (known as “synthetic viability”) (Jeggo *et al.*, 2016). In addition, defects in a particular pathway can make the cell reliant on an alternative pathway; defects in homologous recombination, for example, confer reliance on the non-homologous end joining pathway for the repair of DSBs. While this presents a therapeutic opportunity (see below), it can also promote genomic instability, as NHEJ is inherently more error-prone than HR (Lord and

Ashworth, 2012).

On the other hand, defects in specific pathways, combined with the reliance on remaining pathways for survival, provides therapeutic opportunities. Cancer cells with a specific DNA repair pathway defect may be uniquely susceptible to types of DNA damage normally repaired by this mechanism. An example is the sensitivity of cancers with defects in homologous recombination (for example, BRCA1 mutation or silencing) to cross-linking agents such as cisplatin and to DSBs induced by ionising radiation and topoisomerase poisons (Curtin, 2012). Pharmacological inhibition of upregulated repair pathways can also assist in the shift from survival to cell death (“tipping the balance”, as shown in Figure 1.18), thereby increasing the sensitivity of cancer cells to DNA-damaging treatments. In cases where there is a specific DNA repair pathway deficiency, inhibitors of the remaining intact pathway can cause cell death in the absence of exogenous DNA damage, while causing minimal harm to normal cells that do not have the DNA repair defect. This is known as “synthetic lethality” and an excellent example is the use of PARP inhibitor in patients with HR-defective cancers. These concepts provide a rationale for the development of treatments such as CHK1, DNA-PKcs and PARP inhibitors (Velic *et al.*, 2015).

The recognition that DDR proteins also frequently function as transcriptional regulators provides an additional perspective on the potential roles of these proteins in cancer. In particular, inhibition of these proteins may have consequences not only on DNA repair but also on the transcriptional programs of cancer cells. A detailed understanding of these effects is therefore essential to the effective use of DDR inhibitors in cancer treatment.



**Figure 1.20: The Role of the DNA-Damage Response in Cancer**

The DNA-damage response plays multiple roles in the development and progression of cancer. In normal cells and early cancer cells, DDR proteins are tumour-suppressive, functioning to prevent genomic instability. Mutation or silencing of DDR proteins can therefore promote genomic instability and cancer (for example, in hereditary breast cancer). The DDR also prevents genomic instability induced by oncogenes (replication stress), activating DNA repair as well as cell cycle arrest and senescence/apoptosis pathways; this prevents the propagation of damaged DNA to daughter cells and forms a barrier to cancer progression. However, loss of essential DDR proteins such as p53 causes evasion of cell death/senescence, allowing continued proliferation of genomically altered (and unstable) cells. Paradoxically, however, DDR proteins such as CHK1 and DNA-PKcs are tumour-promoting in the later stages of cancer. Upregulation of DNA repair contributes to avoidance of mitotic catastrophe and other forms of cell death that may arise due to ongoing endogenous DNA damage. Furthermore, upregulation of these proteins decreases the sensitivity of cancer cells to DNA-damaging cancer therapies, which aim to induce catastrophic levels of DNA damage and cell death. However, as well as these challenges,

dysregulation of the DDR provides unique therapeutic opportunities. Cancer cells may have specific susceptibilities to particular types of DNA-damaging therapies due to defects in repair pathways. They may also be reliant on a single intact repair pathway for survival and therefore susceptible to inhibition of this pathway. An additional therapeutic opportunity arising from the dysregulation of the DDR in cancer is the treatment-induced reactivation of proteins promoting cell death or senescence.

## **DNA-PKcs in Cancer**

There have been numerous studies examining DNA-PKcs expression in cancer. This is a complex issue to untangle, with the DNA repair functions of DNA-PKcs influencing aspects of cancer development, progression and response to therapy (in potentially contrasting ways). DNA-PKcs has been proposed to have a tumour suppressor function due to its ability to safeguard genomic integrity but, in contrast, may assist in the survival of cancer cells by protecting them from DNA-damage-induced cell death and contribute to ongoing genomic instability by downregulating HR. In addition, DNA-PKcs also regulates a number of other cellular processes that are highly relevant to cancer, including mitosis, telomere maintenance, metabolism, immunity, the hypoxic response, hormone signalling and transcription. This diversity in DNA-PKcs functions may account for some of the conflicting evidence regarding the role of DNA-PKcs in cancer.

As with other DDR proteins, DNA-PKcs may function as a tumour suppressor in normal cells. Therefore, complete loss or hypomorphic mutations of DNA-PKcs may contribute to cancer susceptibility. Several studies have reported that DNA-PKcs activity (measured by an *in vitro* kinase assay) is lower in peripheral blood lymphocytes from patients with lung, breast or cervical cancer who have not undergone treatment compared to healthy controls (Auckley *et al.*, 2001; Someya *et al.*, 2005). This has been interpreted to reflect a generalised, pre-existing decrease in DNA repair ability, conferring an increased risk of cancer. However, the underlying reasons for this decrease in activity have not been determined and it is unclear if the decrease in activity is a cause or consequence of the disease.

Nonetheless, an interesting finding from these studies is that there is a wide range of DNA-PKcs activity amongst individuals (including healthy controls). One reason for this variation may be genetic polymorphisms in the DNA-PKcs gene, which could alter expression or enzymatic activity. As discussed previously, there are no known cases of complete loss of DNA-PKcs expression or activity in humans, indicating that this germline defect is likely to be embryonic lethal. However, it has been hypothesised that

more subtle effects on the expression or activity of DNA-PKcs may result in low-level DNA damage that escapes checkpoint surveillance, thereby promoting genomic instability and cancer development (Fu *et al.*, 2003). The potential relevance of such polymorphisms to human cancer is demonstrated by murine models. Female BALB/c mice, for example, are unusually sensitive to mammary cancer following treatment with ionising radiation due to two polymorphisms in the coding region of the *DNA-PKcs* (*PRKDC*) gene that reduce DNA-PKcs protein stability (Fabre *et al.*, 2011; Yu *et al.*, 2001). In humans, a number of polymorphisms in *DNA-PKcs* have been examined for their possible association with cancer risk, including rs7003908 (T/G, intronic), rs2213178 (C/T, intronic), rs10109984 (T/C, intronic) and rs7830743 (A/G, exonic). In general, however, the results of these studies have been inconsistent. A recent meta-analysis of 11 studies demonstrated no modification of risk associated with 3 of these polymorphisms (rs2213178 not examined) in breast cancer, bladder cancer and glioma; however, the rs7003908 GG genotype was associated with an increased risk of prostate cancer (Zhang *et al.*, 2013). It is currently unknown how this intronic polymorphism might affect DNA-PKcs expression or function. In addition, the reasons for the tissue-specific alteration in cancer risk are unknown, although it is intriguing to speculate that it may relate to the recently identified transcriptional functions of DNA-PKcs in prostate cancer (Goodwin *et al.*, 2015).

Numerous studies have also examined alterations in DNA-PKcs expression and activity in established cancers and how these alterations relate to survival and treatment outcomes (Table 1.9). DNA-PKcs is frequently upregulated in multiple cancer types, including cervical, liver, lung, oesophageal, pancreatic and prostatic carcinomas. In many cases, high DNA-PKcs expression (either compared to paired normal tissues or within the tumour cohort) is associated with poorer survival outcomes. One exception to this may be gastric carcinoma, in which lack of DNA-PKcs protein expression, occurring in about 20% of cases, is associated with increased stage and lymphatic invasion, and poorer overall survival (Lee *et al.*, 2007; Lee *et al.*, 2005b). In addition, high DNA-PKcs expression in several cancer types, including breast, oesophageal and tonsillar carcinoma, has been correlated with improved outcomes following treatment with radiotherapy; this has been hypothesised to relate to the apoptosis-promoting function of DNA-PKcs in tumours with functional p53 (Friesland *et al.*, 2003; Noguchi *et al.*, 2002; Soderlund Leifler *et al.*, 2010). Conversely, high expression has been associated with an increased risk of recurrence following radiotherapy in prostate cancer (Bouchaert *et al.*, 2012).

For several cancer types, the evidence linking DNA-PKcs expression to outcomes is conflicting, and this is particularly striking in the case of breast cancer (Table 1.9). The largest study of DNA-PKcs expression in breast cancer examined over 1000 cases using immunohistochemistry (IHC) and an additional 2000 cases using microarray data from the METABRIC cohort (Abdel-Fatah *et al.*, 2014; Curtis *et al.*, 2012). In this study, low expression of DNA-PKcs (within the tumour cohort) was associated with increased tumour grade and proliferation, and decreased breast cancer-specific and disease-free survival. When tumours were split by ER status, the association between low expression and poor survival was maintained in ER-positive tumours at the protein level (non-significant association at the mRNA level) and for ER-negative tumours at the mRNA level (non-significant association at the protein level). In contrast, other studies have demonstrated an association between high expression of DNA-PKcs in breast cancer and increased tumour grade (Soderlund Leifler *et al.*, 2010; Sun *et al.*, 2017; Treilleux *et al.*, 2007). High expression of DNA-PKcs was also correlated with chemoresistance and decreased overall survival in one study (Sun *et al.*, 2017) and an improved response to radiotherapy (compared to chemotherapy) in another (Soderlund Leifler *et al.*, 2010), suggesting functions of DNA-PKcs in responses to treatment. One limitation of these studies is that all except one compared DNA-PKcs expression within the tumour cohort (high or low) and did not establish the context of expression in comparison to the normal breast. In the single comparative study, DNA-PKcs protein expression in breast cancer was similar to the normal breast in approximately 25% of cases and decreased compared to normal in the remaining 75% (Treilleux *et al.*, 2007), suggesting that loss of DNA-PKcs expression may contribute to breast cancer development or progression. Taken together, however, these studies indicate that the role of DNA-PKcs in breast cancer has not been clearly defined.

Mutations of DNA-PKcs have also been identified in multiple cancer types (Kan *et al.*, 2010; Lawrence *et al.*, 2014; Wang *et al.*, 2008). Most of these mutations are found at relatively low frequencies, with the possible exception of 936S>C that was identified as a novel recurrent mutation in lung squamous cell carcinoma (Wang *et al.*, 2008). Furthermore, although predictions can be made, the functional consequences of these low-frequency mutations are unknown. Interestingly, one breast cancer sample has been shown to have a deletion of glycine 2113, which was also identified in a human SCID patient (but deemed unlikely to be the SCID-causing mutation) (van der Burg *et al.*, 2009; Wang *et al.*, 2008). DNA-PKcs copy number alterations, primarily copy number gains, have also been identified in several cancer types, including breast cancer (Curtis *et al.*, 2012; Shah *et al.*, 2012). However, a challenge in cancer studies

is distinguishing driver mutations, which are responsible for disease development and progression, on a background of genomic instability that also generates many passenger mutations. Studies that use various techniques to overcome this challenge have not identified DNA-PKcs as a driver gene in cancer (Cancer Genome Atlas Research Network, 2012c; Fleuren *et al.*, 2016; Lawrence *et al.*, 2014; Nik-Zainal *et al.*, 2016). An alternative approach to identifying driver genes is to examine driver networks, by integrating gene expression and copy number information with signalling, protein-protein interaction and transcriptional networks. This approach identified DNA-PKcs as a member of driver networks in both ER-positive and triple-negative breast cancers (Dutta *et al.*, 2012).

Additional alterations in DNA-PKcs that have been found in cancer include novel fusion transcripts and changes in splicing. In endometrial cancer, 3/122 tumour samples (2.5%) expressed a carboxypeptidase Q (*CPQ*)-DNA-PKcs fusion transcript that retained the kinase-coding region. However, knockdown of the fusion transcript in cultured cell lines had no effect on cell growth, suggesting that this fusion is likely to be a passenger alteration associated with amplification of the 8q region (Tamura *et al.*, 2015). In breast cancer, DNA-PKcs has been identified as a differentially spliced gene in TNBC, non-TNBC and HER2-positive tumours compared to the normal breast (Eswaran *et al.*, 2013). The functional significance of these changes is unknown, as there have been no previous studies on the splice variants of the very large 86-exon DNA-PKcs (*PRKDC*) gene.



**Table 1.9: DNA-PKcs expression and activity in clinical cancer studies**

Cancer Type	Assays	Samples	DNA-PKcs Expression/Activity	Conclusions	References
Breast carcinoma	IHC	1161	High expression (H-score >260) in 65% of cases, low expression in remaining 35% of cases	Low expression associated with increased grade, mitotic index, tumour de-differentiation (less tubule formation), and decreased breast cancer-specific survival (BCSS) and disease-free survival (DFS). ER-neg and AR-neg tumours more likely to have low expression.	(Abdel-Fatah <i>et al.</i> , 2014)
		Subset of 835 ER-pos tumours	High expression in 67% of cases, low in 33%	Low expression associated with increased grade, mitotic index, tumour de-differentiation, and decreased BCSS and DFS.	
		Subset of 311 ER-neg tumours	High expression in 61% of cases, low in 39%	Low expression associated with higher mitotic index, no association with survival.	
	Microarray	1929 (METABRIC cohort)	High expression (cut-off not specified) in 34% of cases, low in 66%	ER-pos: low expression showed a trend towards decreased BCSS ER-neg: low expression associated with decreased BCSS, even in subset of patients receiving chemotherapy	(Abdel-Fatah <i>et al.</i> , 2014)
	IHC	224	High expression (>75% of tumour cells positive) in 43% of cases	High expression associated with increased grade and tumour size but decreased risk of local recurrence in patients treated with radiotherapy (compared to those treated with chemotherapy)	(Soderlund Leifler <i>et al.</i> , 2010)
	qPCR	59 (paired)*	Increased expression in tumours	High expression associated with increased grade, lymph node metastasis and chemoresistance, and decreased overall survival (regardless of whether chemotherapy was received)	(Sun <i>et al.</i> , 2017)
	IHC	92 patients and 8 normal breast controls	High expression in 26% of tumours (level comparable to normal tissue), low-intermediate expression in 74%	High expression associated with increased grade	(Treilleux <i>et al.</i> , 2007)
Cervical carcinoma	IHC	109	Increased and heterogeneous expression in tumours	No association with complete response to radiotherapy, tumour grade, subtype or survival	(Beskow <i>et al.</i> , 2006)
	IHC	22 (paired initial biopsy and post-RT surgical samples)	Increased expression in post-RT residual tumours compared to initial biopsy, positive correlation between expression pre- and post-RT	Increased expression may contribute to radiotherapy resistance	(Beskow <i>et al.</i> , 2009)
Chronic lymphocytic leukaemia (CLL)	WB, qPCR	54	Higher in del(17p) patients, good correlation between mRNA and protein expression	High mRNA expression (above median) associated with shorter TFI and overall survival (OS)	(Elliott <i>et al.</i> , 2011)
	WB, kinase assay	54	Higher in del(17p) and del(11q) patients (markers of poor prognosis), expression correlated with activity	High expression (above median) associated with shorter time from diagnosis to chemotherapy (treatment-free interval, TFI)	(Willmore <i>et al.</i> , 2008)
CLL, ALL, high-grade lymphoma, myeloma	IHC, WB	61 patients and 5 normal controls	High expression in ALL, high-grade lymphoma and multiple myeloma, lower expression in CLL	Increased expression associated with less differentiated cancers (with the exception of multiple myeloma)	(Holgersson <i>et al.</i> , 2004)
Colorectal carcinoma	IHC	11 (paired), 8 adenomas and	Trend towards decreased expression in adenomas and carcinomas (not statistically	NA	(Rigas <i>et al.</i> , 2001)

		23 normal colon controls	significant), cytoplasmic expression seen in normal colonic cells		
	WB, qPCR	50 (paired)	Increased expression in tumours	High expression (above median in tumour group) associated with decreased overall survival	(Sun <i>et al.</i> , 2016)
Gastric carcinoma	Microarray (expression and copy number)	25 (paired)	Upregulated (at least 2x) in 64% of tumours, associated with frequent copy number gain of chromosome 8q11.21 ( <i>PRKDC</i> locus)	NA	(Cheng <i>et al.</i> , 2012)
	IHC	279	No expression in 23% of cases	Lack of expression associated with increased stage, increased lymph node invasion/metastasis and decreased overall survival	(Lee <i>et al.</i> , 2005b)
	IHC, mutation analysis	801 (normal, gastritis, adenoma, carcinoma) Ca = 564	No expression in normal superficial epithelium, increased expression in gastritis and adenoma. No expression in 20% of gastric carcinomas.	Lack of expression associated with increased stage, increased lymph node invasion/metastasis, increased neutrophil infiltration and decreased overall survival. DNA-PKcs-negative cancers also more likely to have microsatellite instability phenotype and frameshift mutation of DNA-PKcs poly(A) <sub>10</sub> tract.	(Lee <i>et al.</i> , 2007)
Glioma	Kinase assay	36	Wide range of activity in tumours	Increased activity associated with increased tumour grade and decreased <i>in vitro</i> sensitivity to cisplatin	(Shao <i>et al.</i> , 2008)
Hepatocellular carcinoma	Microarray	132 (normal, cirrhotic, dysplastic, HCC) Ca = 91	Increased expression (2.4x) in tumours compared to non-cancerous liver, copy number gains in 55%	No association with survival	(Cornell <i>et al.</i> , 2015)
	IHC	45 (paired)	Increased expression and phosphorylation in tumours, no significant correlation between expression and phosphorylation	High total expression associated with increased grade and decreased time to radiological progression after treatment with doxorubicin TACE, no correlation with overall survival. High expression of phosphorylated DNA-PKcs (S2056) associated with increased grade and decreased overall survival.	(Cornell <i>et al.</i> , 2015)
	IHC, WB, qPCR, kinase assay	62 (paired) and 6 normal liver controls	Increased expression, phosphorylation and activity in tumours, expression correlated with activity	High expression and high activity associated with decreased overall survival. Activity positively correlated with measures of proliferation, genomic instability and microvessel density, and inversely correlated with apoptosis.	(Evert <i>et al.</i> , 2013)
Lung carcinoma (NSCLC)	qPCR	140 (paired)	Increased expression in tumours	High tumour:normal ratio (above median) associated with decreased overall survival, particularly for adenocarcinoma, younger patients, females and light smokers (vs heavy smokers). No association with survival when expression levels in tumours alone, rather than T:N ratios, were analysed.	(Xing <i>et al.</i> , 2008)
Nasopharyngeal carcinoma	IHC	66	High expression (>50% positive cells) in 70% of cases	No association with disease-free or metastasis-free survival	(Lee <i>et al.</i> , 2005c)

Oesophageal carcinoma	IHC	67	High expression (>30% positive cells) in 54% of cases	High expression associated with better response to chemo-radiotherapy, no association with tumour size, stage or grade.	(Noguchi <i>et al.</i> , 2002)
	IHC, WB, kinase assay	13 (paired)	Increased (heterogeneous) expression and increased activity in tumours, expression correlated with activity	NA	(Tonotsuka <i>et al.</i> , 2006)
Oral squamous cell carcinoma	IHC	42 (paired initial biopsy and post-RT surgical samples)	Increased expression in post-RT residual tumours (30 cases) compared to initial biopsy	No association of expression in initial biopsy with response to radiotherapy	(Shintani <i>et al.</i> , 2003)
Ovarian carcinoma	IHC	190	High expression in 71% of cases	High expression associated with serous cystadenocarcinoma subtype, increased stage and grade, and decreased progression-free and cancer-specific survival	(Abdel-Fatah <i>et al.</i> , 2014)
	qPCR	156	Expressed in tumours	High expression associated with decreased cancer-specific survival	(Abdel-Fatah <i>et al.</i> , 2014)
Pancreatic carcinoma	WB	3 (paired)	Increased expression in tumours	NA	(Hu <i>et al.</i> , 2014)
Prostate carcinoma	IHC	132	Positive nuclear DNA-PKcs expression in 49% of cases	Expression associated with increased risk of biochemical recurrence following external beam radiotherapy	(Bouchaert <i>et al.</i> , 2012)
	Microarray	232	Expressed in tumours	High expression (>80 <sup>th</sup> percentile) associated with increased risk of biochemical recurrence, increased risk of metastases and decreased disease-specific and overall survival	(Goodwin <i>et al.</i> , 2015)
	MS	11 (benign or local) and 16 (metastases)	Increased DNA-PKcs phosphorylation in metastases	NA	(Goodwin <i>et al.</i> , 2015)
	IHC	146	Positive nuclear DNA-PKcs expression in 51% of cases	Expression associated with increased risk of biochemical recurrence following brachy-therapy	(Molina <i>et al.</i> , 2016)
	IHC, WB, qPCR	15 (paired BPH and cancer)	Increased expression in tumours	Increased expression associated with higher stage and grade, increased metastases and decreased overall survival	(Zhang <i>et al.</i> , 2017)
Tonsillar squamous cell carcinoma	IHC	79	High expression (>75% of tumour cells positive) in 87% of cases	High expression associated with increased overall survival, no association with tumour grade, size or complete response to radiotherapy (at 1 month post-RT)	(Friesland <i>et al.</i> , 2003)

**Table 1.9.** ALL, acute lymphoblastic leukaemia; AR-neg, androgen receptor-negative; BPH, benign prostatic hyperplasia; Ca, cancer samples; ER-neg/pos, oestrogen receptor-negative/positive; IHC, immunohistochemistry; MS, mass spectrometry; NSCLC, non-small cell lung cancer; qPCR, quantitative polymerase chain reaction; RT, radiotherapy; TACE, transarterial chemoembolisation; WB, western blotting. \* Paired refers to paired normal tissue and tumour samples taken from the same patient.

## DNA-PKcs in Breast Cancer

There is significant evidence suggesting that DNA-PKcs may have important functions in breast cancer. Firstly, as discussed in the previous section, DNA-PKcs expression is altered in breast cancer and is associated with treatment and survival outcomes. The conflicting results of these studies, however, reveal that this is not a straightforward relationship. One possible reason for this may be that the activity of DNA-PKcs, which is known to be regulated by multiple post-translational modifications, is not adequately represented by measures of mRNA or protein amounts. Another possibility is that functions of DNA-PKcs in addition to DNA repair, particularly transcriptional functions, may contribute to breast cancer development and/or progression. Therefore, the molecular subtypes, defined by their unique transcriptomes, may be differentially regulated by DNA-PKcs activity. This relationship could be further explored using subtype-specific analyses of DNA-PKcs expression and activation. DNA-PKcs phosphorylation (a surrogate marker for activation), for example, has been shown to be significantly increased in the basal-like subtype of breast cancer (Mertins *et al.*, 2016).

Secondly, DNA-PKcs is an important regulator of transcription by ligand-bound hormone receptors. In the ER-positive breast cancer cell line MCF7, DNA-PKcs phosphorylates ER at serine 118, resulting in increased ER transcriptional activity and stabilisation of ER protein levels (Medunjanin *et al.*, 2010b). DNA-PKcs-mediated phosphorylation also guides co-factor dynamics at ER target gene regulatory elements, resulting in the activation of ER co-activators (including SRC3 and MED1) and the dismissal of transient ER co-repressors (such as nuclear receptor interacting protein 1, NRIP1) (Foulds *et al.*, 2013). The dynamic recruitment of DNA-PKcs to ER-bound enhancers is evident at the genome-wide level, as demonstrated by ChIP-seq (Liu *et al.*, 2014). In turn, activated ER positively regulates DNA-PKcs expression and activity, thereby establishing a regulatory circuit between these two proteins (Medunjanin *et al.*, 2010a). A similar relationship has been identified in prostate cancer cells for AR and DNA-PKcs (Goodwin *et al.*, 2015; Goodwin *et al.*, 2013). Furthermore, DNA-PKcs has been shown to regulate cancer-associated transcriptional networks through recruitment to regulatory loci by DNA-bound transcription factors including AR and SP1, contributing to prostate cancer progression and metastasis (Goodwin *et al.*, 2015). These studies clearly point to an important role of DNA-PKcs as a transcriptional modulator in hormone-dependent cancers.

Another intriguing finding from prostate cancer is that DNA-PKcs is essential for the transcriptional activity of ETS factor fusion proteins such as TMPRSS2-ERG. These

fusion events, occurring in approximately 50% of prostate cancers, place the *ERG* or *ETV1* gene under the control of the androgen-regulated promoter and 5' untranslated region of transmembrane protease serine 2 (*TMPRSS2*). This results in AR-driven overexpression of the ETS protein and abnormal transcription of cancer-associated genes (reviewed in Feng *et al.*, 2014). DNA-PKcs, along with PARP1, is recruited by DNA-bound ETS proteins and is required for ETS-driven transcription in prostate cancer cells, while depletion or pharmacological inhibition of DNA-PKcs or PARP1 results in growth inhibition of ETS-positive (but not ETS-negative) prostate cancer cell lines (Brenner *et al.*, 2011). This study suggests that cancers driven by oncogenic ETS factors may be reliant on DNA-PKcs activity for growth and has potential implications for the subset of breast cancers that demonstrate increased ELF5 expression.

DNA-PKcs has also been implicated in the response to the selective oestrogen receptor modulator tamoxifen. In a high-throughput screen using MCF7 breast cancer cells, siRNA-mediated depletion of DNA-PKcs enhanced the sensitivity of cells to tamoxifen (Iorns *et al.*, 2009). Although no validation experiments were performed, this observation raises some intriguing possibilities regarding the role of DNA-PKcs in anti-oestrogen responsiveness, including regulation of tamoxifen-ER transcriptional activity and potential cross-talk with epidermal growth factor (EGFR) signalling. EGFR is known to interact with DNA-PKcs in the nucleus following ionising radiation, enhancing DNA-PKcs DNA repair activity, and overexpression of EGFR is a mechanism of anti-oestrogen resistance (Dittmann *et al.*, 2005; Hiscox, 2009). However, the functional outcomes of the EGFR/DNA-PKcs interaction in areas other than DNA repair have not been explored.

Combined, the above studies point towards important roles of DNA-PKcs in transcriptional regulation and survival signalling in breast cancer.

## Chapter 2: Materials and Methods

### Cell-Based Methods

#### Stable cell line generation

ELF5 Isoforms 1 and 3 were tagged with C-terminal V5 (and short linker sequence), cloned into the pHUSH-ProEx vector (Gray *et al.*, 2007) and used as a retrovirus. T47D-EcoR and MDA-MB-231-EcoR cells, stably expressing ecotropic receptor, were infected with pHUSH-ELF5 retrovirus and selected using puromycin. To generate clonal cell lines, stable cell line pools were plated at low density in 96-well plates. ELF5 Isoform 2 cell lines (pools and clones) were previously created in an identical manner (Kalyuga *et al.*, 2012). A diagram of the pHUSH-ELF5 vector construct is shown in Figure 3.13.

#### Cell lines and treatments

A list of all cell lines and abbreviations used throughout this thesis is shown in Table 2.1. All cell lines were maintained in Gibco RPMI 1640 medium (Thermo Fisher Scientific, Waltham, Massachusetts, USA) supplemented with 10% foetal bovine serum (Thermo Fisher) or 10% Tetracycline-free foetal bovine serum (Clontech, Mountain View, California, USA). Medium was also supplemented with 10ug/mL insulin (Novo Nordisk, Bagsvaerd, Denmark). Puromycin (Sigma-Aldrich, St Louis, Missouri, USA) was added at a concentration of 1ug/mL to maintain selection pressure. Doxycycline (Dox, Sigma-Aldrich) was added at a concentration of 0.1ug/mL daily to induce ELF5 protein expression (vehicle control = water). For experiments involving hormone deprivation, cells were cultured for 72 hours in phenol red-free RPMI (Thermo Fisher) supplemented with insulin and 10% charcoal-stripped FBS. Oestradiol (Sigma-Aldrich) was added to hormone-deprived cells at a concentration of 100nM (vehicle control = 100% ethanol), followed by collection at specified times. All cell cultures were maintained in a 37°C, 5% CO<sub>2</sub> humidified incubator.

**Table 2.1: Cell lines**

Cell line (full name)	Cell line (abbreviated)	Description	Variants	Molecular features of parental line*
MCF7-EcoR-pHUSH-ProEx-ELF5-Isoform2-V5	MCF7-ELF5-Iso2-V5 (or MCF7-ELF5-V5)	MCF7 cells modified with ecotropic receptor and stably infected with pHUSH-ProEx retroviral vector containing ELF5 Isoform 2 tagged with V5 epitope	Pooled cell line only (no clonal selection)	Tumour source: metastasis (pleural effusion) Molecular subtype: luminal ER/PR: Positive HER2: Negative Known mutations: CDKN2A, PIK3CA p53 status: wild-type
MCF7-EcoR-pHUSH-ProEx-empty	MCF7-pHUSH-empty	MCF7 cells modified with ecotropic receptor and stably infected with pHUSH-ProEx retroviral vector with no gene insert	Pooled cell line only (no clonal selection)	
MDA-MB-231-EcoR-pHUSH-ProEx-ELF5-Isoform2-V5	MM231-ELF5-Iso2-V5	MDA-MB-231 cells modified with ecotropic receptor and stably infected with pHUSH-ProEx retroviral vector containing ELF5 Isoform 2 tagged with V5 epitope	Pooled cell line Clonal cell lines (numbered 1, 6, 7)	Tumour source: metastasis (pleural effusion) Molecular subtype: claudin-low Known mutations: BRAF, CDKN2A, KRAS, NF2, p53 ER/PR: Negative HER2: Negative p53 status: mutated
MDA-MB-231-EcoR-pHUSH-ProEx-ELF5-Isoform3-V5	MM231-ELF5-Iso3-V5	As above for ELF5 Isoform 3	Pooled cell line Clonal cell lines (numbered 2, 7, 20, 22)	
MDA-MB-231-EcoR-pHUSH-ProEx-empty	MM231-pHUSH-empty	MDA-MB-231 cells modified with ecotropic receptor and stably infected with pHUSH-ProEx retroviral vector with no gene insert	Pooled cell line only (no clonal selection)	
T47D-EcoR-pHUSH-ProEx-ELF5-Isoform1-V5	T47D-ELF5-Iso1-V5	MDA-MB-231 cells modified with ecotropic receptor and stably infected with pHUSH-ProEx retroviral vector containing ELF5 Isoform 1 tagged with V5 epitope	Pooled cell line Clonal cell lines (numbered 2, 9, 10, 16)	Tumour source: metastasis (pleural effusion) Molecular subtype: luminal ER/PR: Positive HER2: Negative Known mutations: p53, PIK3CA p53 status: mutated
T47D-EcoR-pHUSH-ProEx-ELF5-Isoform2-V5	T47D-ELF5-Iso2-V5	As above for ELF5 Isoform 2	Pooled cell line Clonal cell lines (numbered 8, 9, 13)	
T47D-EcoR-pHUSH-ProEx-ELF5-Isoform3-V5	T47D-ELF5-Iso3-V5	As above for ELF5 Isoform 3	Pooled cell line Clonal cell lines (numbered 10, 11, 20, 26)	
T47D-EcoR-pHUSH-ProEx-empty	T47D-pHUSH-empty	T47D cells modified with ecotropic receptor and stably infected with pHUSH-ProEx retroviral vector with no gene insert	Pooled cell line only (no clonal selection)	

BRAF, B-Raf proto-oncogene, serine/threonine kinase; CDKN2A, cyclin dependent kinase inhibitor 2A; ER, oestrogen receptor; HER2, erb-b2 receptor tyrosine kinase 2; KRAS, KRAS proto-oncogene, GTPase; NF2, neurofibromin 2; p53, tumour protein p53; PIK3CA, phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit alpha; PR, progesterone receptor. \*Molecular information: (American Type Culture Collection, 2015; Neve *et al.*, 2006; Prat *et al.*, 2010).

### **Cell number assessment**

Cell number was quantified as an experimental end-point using either automated cell counts or a spectrophotometric assay. Cell counts were performed on the Countess automated cell counter (Thermo Fisher), using 0.4% trypan blue to exclude dead cells. For the cell number assay, cells grown in a 6-well plate were incubated with 16% trichloroacetic acid, followed by staining with 10% Diff-Quick II solution (Lab Aids, North Narrabeen, NSW, Australia). Acetic acid (10%) was added to the dried plates and 100 $\mu$ L of solution from each well was added to a 96-well plate, which was read at 595nm. In Figure 3.13, absorbance readings were transformed to natural logarithms and values from 3 wells (single experiment) were averaged for each time point; the  $\pm$  doxycycline slopes for each cell line were compared using Prism linear regression analysis (GraphPad, La Jolla, California, USA). In Figure 5.14, absorbance from 3 or 4 biological replicates were read simultaneously; differences between experimental conditions were then compared within each cell line using analysis of variance (ANOVA) followed by Tukey's multiple comparisons test (GraphPad Prism).

### **Transient retroviral infection**

ELF5 Isoform 3 was tagged with C-terminal haemagglutinin (HA) tag, cloned into the pQCXIH vector (Clontech) and used as a retrovirus. MDA-MB-231-EcoR-pHUSH-ELF5-Isoform2-V5 Clone 7 cells were infected with ELF5-Isoform3-HA or empty vector virus diluted 1:4. No pQCXIH selection pressure was applied.

### **Immunofluorescence**

Cells (see above) were infected with pQCXIH retrovirus in 8-well Lab-Tek II chamber slides (Thermo Fisher) and allowed to recover for 24 hours. Doxycycline or vehicle treatment (lasting 24 hours) was then commenced. Cells were fixed with 4% formaldehyde (Thermo Fisher) diluted in PHEM buffer (60mM PIPES, 25mM HEPES, 1mM EGTA, 2mM MgCl<sub>2</sub>, pH 6.9), permeabilised with 0.5% Triton X-100, and blocked with 10% donkey serum (Jackson ImmunoResearch Laboratories, West Grove, Pennsylvania, USA). Primary antibody incubation was performed overnight at 4°C. Secondary antibodies were added at 1:200 and coverslips applied using Duolink In Situ Mounting Medium with DAPI (Olink Bioscience, Uppsala, Sweden / Sigma-Aldrich). Imaging was performed on a Leica DM5500 microscope (Leica Microsystems, Wetzlar, Germany). Antibodies (in 10% donkey serum/PHEM solution): anti-V5 1:200 (sc-58052, Santa Cruz Biotechnology (SCBT), Dallas, Texas, USA), anti-HA 1:800 (#3724, Cell Signaling Technology (CST), Danvers, Massachusetts, USA), donkey anti-mouse



AlexaFluor 647 and donkey anti-rabbit AlexaFluor 555 conjugates 1:200 (Thermo Fisher).

### **siRNA transfection**

ON-TARGETplus human *PRKDC* SMART pool siRNA (Dharmacon, Lafayette, Colorado, USA) was resuspended in nuclease-free sterile water at a concentration of 100uM, and stored as single-use 5uL aliquots at -70°C. ON-TARGETplus non-targeting siRNA #1 (Dharmacon) was used as a control. All transfections were performed using Lipofectamine RNAiMAX transfection reagent (Thermo Fisher). The optimal transfection parameters were established as 5nM siRNA (0.12uL) and 2.5uL Lipofectamine per well (6-well plate) in a total volume of 2.4mL. The siRNA and Lipofectamine mixture was diluted in Opti-MEM (Thermo Fisher) and incubated at room temperature for approximately 20 minutes. 400uL of siRNA/Lipofectamine mixture was then added to each well (or Opti-MEM only for the untransfected control), followed by 2.0mL of cells suspended in normal medium. Cell numbers for 6-well plates (Corning Life Sciences, Tewksbury, Massachusetts, USA) were 80,000 cells/well (MCF7 lines), 150,000 cells/well (T47D lines), or 40,000 cells/well (MDA-MB-231 lines). No antibiotics, including puromycin, were added during transfection. After 24 hours, the medium was changed and puromycin was commenced to maintain doxycycline-induced ELF5 expression. Doxycycline treatment was started on day 2 and cells were collected on day 4 (see Timeline in Figure 5.12D).

### **Treatment with ionising radiation**

MCF7-pHUSH-empty cells in 10cm plates were treated with 0 Gy (control), 2 Gy or 10 Gy ionising radiation, and collected 40 minutes post-treatment.

### **Proximity ligation assays (PLAs)**

MCF7-ELF5-Isoform2-V5 or MCF7-pHUSH-empty cells were seeded on glass coverslips in 12-well plates (Corning) and treated with doxycycline or vehicle for 48 hours. Three biological replicates were performed. Coverslips were washed x 2 in room temperature Dulbecco's phosphate-buffered saline (PBS, Thermo Fisher), fixed for 10 minutes with 4% PFA diluted in PHEM buffer (see above), permeabilised for 10 minutes with 0.5% Triton X-100, and blocked with 10% donkey serum for 2 hours at 37°C (Jackson ImmunoResearch). Primary antibody incubation was conducted overnight at 4°C using the following antibodies (or combinations of antibodies) diluted in 10% donkey serum/PHEM solution: anti-V5 1:1000 (#13202, CST), anti-DNA-PKcs

1:50 (#12311, CST), anti-DNA-PKcs 1:1000 (MS-423-P1, Thermo Fisher). Isotype control antibodies were diluted to ensure equivalent amounts of antibody to matched primary: Mouse IgG1 (Dako X0931, Agilent, Santa Clara, California, USA), Mouse IgG2a (Dako X0943, Agilent), Rabbit IgG (NB810-56910, Novus Biologicals, Littleton, Colorado, USA). On day 2, the standard Duolink protocol was followed (Sigma-Aldrich, protocol summarised in Figure 5.6). Briefly, coverslips were incubated for 1 hour at 37°C with Duolink green minus and plus probes (rabbit and mouse), according to the species of primary antibody/antibodies used. This was followed by incubation with ligation solution (30 mins at 37°C) and amplification solution (100 mins at 37°C). Total reaction volumes were 40uL per coverslip. 4 x washes in Duolink wash buffer A (0.01M Tris, 0.15M NaCl, 0.05% Tween-20) were carried out between each step and the final 3 x washes in Duolink wash buffer B (0.2M Tris and 0.1M NaCl, undiluted or 1:100 as per protocol). Coverslips were allowed to dry in the dark and then mounted on glass slides with Duolink In Situ Mounting Medium with DAPI. Slides were stored at 4°C for a period of up to 3 weeks until imaging.

## **Image-Based Methods**

### **PLA image acquisition**

PLA coverslips were imaged 8-18 days after completion of the Duolink staining protocol on a Leica DM5500 microscope (Leica Microsystems). To ensure unbiased image acquisition, groups of cells to be imaged were identified by horizontal movement across the coverslip for a total of 4 rows (separated by 2mm) using the DAPI (nuclear) A4 filter cube. The PLA signals were briefly viewed (using the L5 green Leica filter cube) to ensure they were in focus and the image was acquired using pre-defined exposure times. In this way, the cells for PLA quantification were selected without knowledge of the PLA signal level. Approximately 15-120 images were taken per coverslip; the total number of cells (nuclei) for the combined replicates are shown in Figure 5.10. Due to technical issues, only two (of three) experimental replicates were imaged for the V5 and DNA-PKcs CST antibody combination and controls.

### **PLA image analysis**

PLA images were analysed using a FIJI (Schindelin *et al.*, 2012) macro created by Andrew Law in this laboratory (Law *et al.*, 2017b). Nuclear images were modified with 'Enhance Contrast' at a saturation value of 0.4, 'Subtract Background' with a rolling ball radius of 100, and then converted to an 8-bit greyscale image, and thresholded. They

were then processed with 'Watershed' and 'Analyse Particles' to select for and create a mask image of the nuclei. Foci (PLA signals) were then selected using the 'Find Maxima' function and a single point mask was created from the foci selection. The mask images of the foci and nuclei were added together with the 'Image Calculator' function, and non-nuclear signals defined. Nuclear signals were then calculated by subtracting non-nuclear signals from the total signal number. Chi-square analysis of the signal distribution was performed using GraphPad Prism.

## **Protein-Based Methods**

### **Rapid Immunoprecipitation of Endogenous Protein (RIME)**

MCF7-ELF5-Isoform2-V5 cells were grown in multiple 15cm plates (Corning), treated with doxycycline for 72 hours, then cross-linked using 1% methanol-free formaldehyde (Thermo Fisher) diluted in serum-free RPMI medium. After 10 minutes, formaldehyde was quenched with 0.2M glycine. Plates were then placed on ice and washed x 2 with cold PBS. 1mL of PBS containing calcium and magnesium salts (Thermo Fisher) and Complete EDTA-free Protease Inhibitor Cocktail (Roche, Basel, Switzerland) was added to the cells, which were then collected using a cell scraper. Cells were pooled and pellets containing approximately 50 million or 100 million cells were stored at -70°C.

RIME was performed using cross-linked cell pellets according to the previously published protocol with some modifications (Mohammed *et al.*, 2013). For RIME experiment 1, cross-linked cell pellets were shipped to the laboratory of Dr Jason Carroll (Cancer Research UK Cambridge Institute), and ELF5-V5 RIME was performed as per the published protocol using a 1:1 mix of anti-ELF5 and anti-V5 antibodies (see below). RIME experiments 2-5 were performed in collaboration with Drs Mark Molloy and Christoph Krisp at the Australian Proteome Analysis Facility (APAF, Macquarie University, Sydney). Cell pellets x 2 (100 million cells for experiments 2-4 and 50 million cells for experiment 5) were thawed on ice and resuspended in 10mL of LB1 buffer (50mM HEPES-KOH pH 7.5, 140mM NaCl, 1mM EDTA, 10% glycerol, 0.5% Igpal CA-630, and 0.25% Triton X-100) with protease inhibitors (Roche Complete Protease Inhibitor cocktail, with added Mg132, sodium vanadate and dithiothreitol). The cells were rotated at 4°C for 30-60 minutes, pelleted, and the supernatant (cytoplasmic fraction) removed. Cells were resuspended in 10mL of LB2 buffer (10mM Tris-HCl pH 8.0, 200mM NaCl, 1mM EDTA, 0.5mM EGTA, protease inhibitors), rotated at 4°C for 30 minutes, and pelleted. Finally, cells were resuspended in 6mL of LB3 buffer (10mM

Tris-HCl pH 8.0, 100mM NaCl, 1mM EDTA, 0.5mM EGTA, 0.1% sodium deoxycholate, 0.5% N-lauroylsarcosine, protease inhibitors) and the two samples pooled.

The pooled sample was then aliquoted into microcentrifuge tubes (300uL per tube) and sonicated using a probe sonicator (4 x 10 second cycles on ice). Triton X-100 (10%) diluted in LB3 was added to each sample tube (30uL), and the lysates were centrifuged for at 20,000 rcf for 10 minutes at 4°C to purify debris. The supernatants from each tube were then re-pooled. A small volume (~50uL) was put aside for subsequent western blot analysis of input protein, as well as reverse cross-linking, DNA purification, and agarose gel analysis of fragment sizes. The sample was then divided into 2 x 15mL tubes and incubated on a rotator overnight at 4°C with Pierce protein A/G magnetic beads (Thermo Fisher). The beads (110uL per immunoprecipitation) were pre-bound with a 1:1 combination of either 10ug anti-V5 (R-960-25, Thermo Fisher) and 10ug anti-ELF5 N-20 (sc-9645, SCBT) primary antibodies, or equivalent amounts of mouse IgG2a (Dako X0943, Agilent) and goat IgG (sc-2028, SCBT) isotype control antibodies. Following the overnight incubation, the beads were washed 5 times in 1mL of RIPA buffer, with approximately 10% of the sample reserved for western blot analysis after the final wash. The remaining beads were washed twice in 100mM ammonium hydrogen carbonate (AMBIC) solution, with the sample transferred to a new microcentrifuge tube after the first AMBIC wash. The beads were then resuspended in 50uL AMBIC solution and delivered to APAF, where the remainder of the protocol was performed by Christoph Krisp.

Beads were diluted with 100 µL of 100mM triethylammonium bicarbonate and 1% sodium deoxycholate and boiled at 95°C for 5 minutes. After disulfate bond reduction with 10 mM dithiothreitol (30 minutes at 60°C) and cysteine alkylation with 20 mM iodoacetamide (30 minutes at 37°C in the dark), each sample was digested with trypsin overnight at 37°C (about 20:1 protein to enzyme). Supernatant was removed from the beads and transferred to a new tube. Samples were acidified with formic acid (1% final concentration) to quench digestion and precipitate sodium deoxycholate. Samples were then spun at 14,000 rcf for 5 min and the supernatant transferred to a new tube. Samples were dried in a vacuum concentrator and resuspended in 20 µL 2% acetonitrile (ACN) and 0.1% formic acid.

Next, 10uL of each sample was injected onto a C18 reversed phase (RP) peptide trap chip (0.5 mm, 200 µm, 300Å ChromXP C18 RP) for purification. Peptides were eluted from the trap and separated on a 15cm chip column (200 µm, 300Å ChromXP C18 RP) using a linear solvent gradient from 5% ACN 0.1% formic acid to 40% ACN, 0.1%

formic acid at 600nL/min over a 60 min period. The LC eluents were subject to positive ion nanoflow electrospray MS analysis in an information dependant acquisition mode (IDA) on a 5600 TripleToF mass spectrometer with Eksigent NanoLC Ultra with cHiPLC system (SCIEX, Framingham, Massachusetts, USA).

In information dependent acquisition (IDA) mode, a TOF-MS survey scan was acquired (m/z 350-1500, 250 ms accumulation time), then the 20 most intense multiply charged ions (2+ - 4+; counts >150cps) in the survey scan were sequentially subjected to MS/MS analysis. MS/MS spectra were accumulated for 100 ms (m/z 100-1500) with rolling collision energy. IDA data were searched against the Human SwissProt data base release April 2014 with ProteinPilot software version 4.2 (SCIEX) using the mascot algorithm. Decoy database search was enabled to allow false discovery rate (FDR) calculation. Proteins accepted to be present in a sample had a FDR < 1%.

### **Co-immunoprecipitations for western blot**

ELF5/V5 co-immunoprecipitations (shown in Figure 5.5) were prepared according to the protocol described above. After the final RIPA buffer wash, approximately 10% of the IP beads were pelleted and stored at -20°C. For western blot analysis, the beads were resuspended in 20uL of NuPAGE Sample Buffer containing 2x Reducing Agent (Thermo Fisher) and incubated at 95°C for 5 minutes. The supernatant was then transferred to a fresh tube and run on a pre-cast 4-12% Bis-Tris polyacrylamide gel (Thermo Fisher).

### **Phosphoprotein purification**

Approximately  $10 \times 10^6$  cells with DNA-PKcs knockdown +/- doxycycline treatment were collected and purified using the Qiagen Phosphoprotein Purification Kit (Qiagen, Hilden, Germany). Cell lysis and phosphoprotein purification were performed according to the standard kit protocol. Phosphoprotein fractions 3 and 4 were pooled and concentrated to a final volume of 30-50uL using NanoSep columns with Omega membrane 10kDa MWCO (Pall Corporation, Port Washington, New York, USA).

### **Western blots**

Western blots were performed according to the following general protocol; more specific information is provided in Tables 2.2 (experimental details) and 2.3 (antibody details). Whole cell lysates were prepared from adherent cells or cell pellets using Normal Lysis Buffer (1.2% HEPES, 1% Triton X-100, 10% glycerol, 0.8% NaCl, 0.03%

MgCl<sub>2</sub>, 0.04% EGTA, 1.0% disodium pyrophosphate, 0.4% NaF), except where otherwise indicated. Roche Complete EDTA-free Protease Inhibitor Cocktail and PhosSTOP phosphatase inhibitors (for phospho-blot samples) were added to the lysis buffer at 1x concentration. Cells were incubated on ice, vortex, and centrifuged at 10,000rpm for 10 minutes at 4°C; supernatant was collected and stored at -70°C. Protein concentration was measured using the Bio-Rad Protein Assay (Bio-Rad Laboratories, Hercules, California, USA). Cell lysate samples were prepared using NuPAGE Sample Buffer and Reducing Agent (Thermo Fisher), heated at 70°C for 10 minutes, and run on pre-cast NuPAGE gels (Thermo Fisher) in MOPS buffer or MES buffer (for selected ELF5 blots). Tris-acetate gels (phospho-DNA-PKcs antibodies, Figure 5.21B) were run in Tris-Acetate buffer (50mM tricine, 50mM tris base, 0.1% SDS). Proteins were transferred to polyvinylidene difluoride (PVDF) membrane at 100V for 1 hour (increased to 2 hours for selected blots examining DNA-PKcs due to its large size). Membranes were cut at specific molecular weights, guided by Precision Plus Protein Dual-Colour Standards (Bio-Rad Laboratories), to facilitate incubations with multiple blocking solutions and primary antibodies. Membranes were blocked for 1-2 hours at room temperature in TBS-tween (10mM Tris base, 150mM NaCl, 0.1% Tween) with either 5% skim milk, 5% bovine serum albumin (BSA, Sigma-Aldrich, phospho-ER and phospho-DNA-PKcs membranes), or 5% donkey serum (Jackson ImmunoResearch, ELF5 N-20 membranes). Primary antibody incubation was performed overnight at 4°C with gentle shaking. Membranes were then incubated with secondary HRP-conjugated antibody diluted 1:2000-1:5000 in blocking solution for 1 hour at room temperature with gentle shaking. TBS-tween washes were performed after each antibody incubation. Proteins were detected using enhanced chemiluminescence solution (Western Lightning Plus, Perkin Elmer, Waltham, Massachusetts, USA) and x-ray film (Fujifilm, Tokyo, Japan).

**Table 2.2: Western blot experiments**

Description	Figures	Gel type	Loading amount	Primary antibodies	Other notes
pHUSH-ELF5 cell lines over-expressing ELF5 Isoforms 1, 2, or 3 (demonstrating different molecular weights)	3.12C	4-12% Bis-Tris, 26-wel	30ug	V5 (CST)	
Panel of breast cancer lines	3.12D	4-12% Bis-Tris, 26-well	Maximum volume (11.7uL) capped at 100ug protein	ELF5, $\beta$ -actin	Breast cancer cell lines classified according to molecular subtype (Prat <i>et al.</i> , 2013). Gel run in MES buffer.
T47D- and MDA-MB-231- ELF5 isoforms clonal cell lines timecourse, blots for ER and related proteins	3.13E 3.13F	4-12% Bis-Tris, 15-well	10ug	TLE1, ER (SCBT), FOXA1 (SCBT), $\beta$ -actin	Multiple gels run (using the same lysates) due to large numbers of proteins; images from the same gel are shown in outlined boxes.
T47D- and MDA-MB-231- ELF5 isoforms clonal cell lines, V5 blots	3.13H	10% Bis-Tris, 10-well	25ug, except for T47D-ELF5-Iso2-V5 clone 8 (65ug)	V5 (SCBT), $\beta$ -actin	
Co-IPs for cross-linked MCF7-ELF5-V5 cells prepared using RIME protocol	5.5	4-12% Bis-Tris, 12-well	Co-IPs: NA (beads) Input and IP washes: 13uL (concentration not measured)	V5 (SCBT), DNA-PKcs (CST)	Beads resuspended in 20uL loading buffer. Membranes initially blotted for V5, then stored in TBS-tween at 4°C and later blotted for DNA-PKcs.
Phosphoprotein purification, MCF7-ELF5-V5 cells treated with doxycycline or vehicle	5.11D	4-12% Bis-Tris Midi Gel, 12+2- well	Total lysate: 20ug Phospho samples: maximum volume of 22.5uL	V5 (CST), $\beta$ -actin	
DNA-PKcs knockdown optimisation experiment 1, MCF7-ELF5-V5 cells	5.12A	4-12% Bis-Tris, 15-well	20ug	DNA-PKcs (CST), $\beta$ -actin	
DNA-PKcs knockdown optimisation experiment 2 with doxycycline treatment, MCF7-ELF5-V5 cells	5.12C	4-12% Bis-Tris, 15-well	15ug	DNA-PK (CST) V5 (CST) $\beta$ -actin	
Phosphoprotein purification, MCF7- (A) or T47D- (B) ELF5-V5 cell lines with DNA-PKcs knockdown	5.13A 5.13B	4-12% Bis-Tris Midi Gel, 12+2- well	Total lysate: 20ug Phosphoprotein samples: 40ug	V5 (CST), FOXA1 (SCBT), $\beta$ -actin	
Phosphoprotein purification, unmodified T47D cells with DNA-PKcs knockdown	5.13C	4-12% Bis-Tris Midi Gel, 12+2- well	Total lysate: 14ug Phosphoprotein samples: 40ug	ELF5, DNA-PKcs (CST), FOXA1 (SCBT), $\beta$ -actin	ELF5 membrane blocked in 5% donkey serum in TBS-tween
MCF7-ELF5-V5 cell lines +/- doxycycline and DNA-PKcs knockdown	5.19A	4-12% Bis-Tris, 15-well	15ug (all replicates)	DNA-PKcs (CST), V5 (CST), ELF5, ER (CST),	Multiple gels run (using the same lysates) due to large numbers of proteins; images from

				phospho-ER S118, AKT, phospho-AKT S473, FOXA1 (Abcam), GATA3 (CST), VTCN1, $\beta$ 1-integrin, E-cadherin, $\beta$ -actin	the same gel are shown in outlined boxes (each with matched loading control). Gels were run for all three experimental replicates and images shown are taken from all replicates.
T47D-ELF5-V5 cell lines +/- doxycycline and DNA-PKcs knockdown	5.19B	4-12% Bis-Tris Midi Gel, 20-well	10ug (all replicates)	DNA-PKcs (CST), V5 (CST), ELF5, ER (CST), phospho-ER S118, AKT, phospho-AKT S473, FOXA1 (SCBT), GATA3 (CST), VTCN1, $\beta$ 1-integrin, E-cadherin, $\beta$ -actin	As above
MDA-MB-231-ELF5-V5 cell lines +/- doxycycline and DNA-PKcs knockdown	5.19C	4-12% Bis-Tris Midi Gel, 20-well	7.5ug (all replicates)	DNA-PKcs (CST), V5 (CST), AKT, phospho-AKT S473, ELF5, $\beta$ 1-integrin, $\beta$ -actin	As above
T47D- and MDA-MB-231- ELF5-V5 clonal cell lines timecourse, blots for DNA-PKcs	5.20	4-12% Bis-Tris, 15-well	10ug	DNA-PKcs (CST), $\beta$ -actin	Same experiment as Figures 3.13E and 3.13F
Phospho-DNA-PKcs antibody optimisation, using irradiated MCF7-pHUSH-empty cells	5.21A	3-8% Tris-Acetate, 10-well	20ug	DNA-PKcs (CST), phospho-DNA-PKcs S2056, phospho-DNA-PKcs T2609, $\beta$ -actin	Gels run in MOPS (not Tris-Acetate) buffer. Phospho-antibody membranes blocked in 5% BSA. Phospho-DNA-PKcs S2056 antibody diluted in TBS/BSA.
MCF7-, T47D- and MDA-MB-231 cell lines -/+ doxycycline, phospho-DNA-PKcs blots	5.21B	3-8% Tris-Acetate Midi Gel, 20-well	MCF7 lines: 40ug T47D lines: 30ug 231 lines: 15ug Irradiated control: 16ug	DNA-PKcs (CST), phospho-DNA-PKcs S2056, $\beta$ -actin	Gels run in Tris-Acetate buffer. Phospho-antibody membrane blocked in 5% BSA. Phospho-DNA-PKcs antibody diluted in 5% milk.
MCF7-ELF5-V5 cells +/- doxycycline and +/- oestradiol treatment	5.22	4-12% Bis-Tris Midi Gel, 20-well	20ug	DNA-PKcs (CST), phospho-DNA-PKcs S2056, ER (CST), pER S118, V5 (CST), AKT, phospho-AKT S473, $\beta$ -actin	Multiple gels run (using the same lysates) due to large number of proteins analysed; images from the same gel shown in outlined boxes (each with matched loading control). Primary antibody incubation was 48 hours.



**Table 2.3: Western blot antibodies**

Antibody target	Product number	Company	Dilution	Diluent	Species and isotype	Blocking agent
AKT	2920	CST	1:2000	TBS/BSA	Mouse IgG1	5% skim milk
Phospho-AKT S473	4060	CST	1:2000	TBS/BSA	Rabbit IgG monoclonal	5% skim milk
$\beta$ 1-integrin or CD29	610467	BD Biosciences	1:1000	TBS/BSA	Mouse IgG1	5% skim milk
$\beta$ -actin	A5441	Sigma-Aldrich	1:20,000	TBS/BSA	Mouse IgG1	5% skim milk
DNA-PKcs (CST)	12311	CST	1:1000	TBS/BSA	Mouse IgG1	5% skim milk
Phospho-DNA-PKcs S2056	Ab18192	Abcam	1:1000	TBS/BSA or 5% skim milk	Rabbit IgG	5% BSA
Phospho-DNA-PKcs T2609	Ab18356	Abcam	1:1000	TBS/BSA	Mouse IgG1	5% BSA
E-cadherin	610182	BD Biosciences	1:1000	TBS/BSA	Mouse IgG2a	5% skim milk
ELF5	Sc-9645	SCBT	1:500-1:1000	5% donkey serum	Goat IgG	5% donkey serum
ER (SCBT)	Sc-8005	SCBT	1:1000	TBS/BSA	Mouse IgG2a	5% skim milk
ER (CST)	8644	CST	1:1000	TBS/BSA	Rabbit IgG monoclonal	5% skim milk
Phospho-ER S118	2517	CST	1:1000	5% skim milk	Mouse IgG2b	5% BSA
FOXA1 (Abcam)	Ab109760	Abcam	1:1000	TBS/BSA	Rabbit IgG monoclonal	5% skim milk
FOXA1 (SCBT)	Sc-101058	SCBT	1:1000	TBS/BSA	Mouse IgG2a	5% skim milk
GATA3 (CST)	5852	CST	1:1000	TBS/BSA	Rabbit IgG monoclonal	5% skim milk
TLE	Ab183742	Abcam	1:1000	TBS/BSA	Rabbit IgG monoclonal	5% skim milk
V5 (CST)	13202	CST	1:1000	TBS/BSA	Rabbit IgG monoclonal	5% skim milk
V5 (SCBT)	Sc-58052	SCBT	1:500-1:1000	TBS/BSA	Mouse IgG2a	5% skim milk
VTCN1 or B7-H4	14572	CST	1:1000	TBS/BSA	Rabbit IgG monoclonal	5% skim milk

BD Biosciences (Franklin Lakes, New Jersey, USA); Abcam (Cambridge, Massachusetts, USA).

## RNA-Based Methods

### End-point PCR

RNA was extracted using the RNeasy Mini Kit with DNase treatment (Qiagen). cDNA was made from 2ug RNA using the Applied Biosystems High Capacity cDNA Reverse Transcription Kit (Thermo Fisher) with RNasin Ribonuclease Inhibitor (Promega, Madison, Wisconsin, USA). PCR reactions were prepared in 50uL total volume using the PCR Reagent System (Thermo Fisher). Each reaction contained cDNA (5uL), 10x buffer (5uL), 10mM dNTPs (1uL, final concentration 200uM), 100uM forward and reverse primers (0.25uL of each, final concentration 0.5uM), Taq DNA polymerase 5U/uL (0.25uL), 50mM magnesium chloride (3uL, final concentration 3mM), and water (35.25uL). Reactions were run on a thermal cycler as follows: 1) Initial denaturation at 94°C for 3 minutes; 2) 25 cycles of denaturation (94°C for 45 seconds), annealing (59°C for 30 seconds), and extension (72°C for 1 minute); 3) Final extension at 72°C for 10 minutes. Amplicons were visualised on a 1% agarose/ethidium bromide gel (Figure 3.12B). *ELF5* Isoform 2/3 primers were designed using NCBI Primer-BLAST (5' to 3'): AGCGCCTGCCTTCTCTTGCC (forward) and CCCCACATCTTTGCCAGGGCTT (reverse).

### Quantitative PCR

RNA was extracted using the RNeasy Mini Kit with DNase treatment (Qiagen) and quantified using the Nanodrop spectrophotometer (Thermo Fisher). cDNA was made using the Applied Biosystems High Capacity cDNA Reverse Transcription Kit (Thermo Fisher) with RNasin Ribonuclease Inhibitor (Promega). All qPCR reactions were run on the Applied Biosystems ABI7900 qPCR machine (Thermo Fisher). Two to three technical replicates were run for each sample, as well as negative controls (no template, no reverse transcriptase, water). Standard curves using a 1:10 dilution series were run for every assay to determine amplification efficiency and relative quantity.

Taqman assays were run using 4.5uL cDNA (diluted 1:5-1:10 in nuclease-free water) and 5.5uL assay (diluted 1:11 in Taqman Gene Expression Mastermix, Thermo Fisher) using standard Taqman cycling conditions. Roche Universal Probe Library (UPL) assays were designed using the online Roche ProbeFinder software. All Roche assays were tested prior to use with a 6-point 1:10 dilution series and assays with poor amplification were not used. Each 10uL Roche qPCR reaction included 0.4uL forward primer (10uM), 0.4uL reverse primer (10uM), 0.1uL UPL probe, 5uL LightCycler 480

Probes Master reaction mix (Roche) and 4.1uL of diluted cDNA. Reactions were run in 384-well plates on the ABI7900 qPCR machine (Life Technologies) using the Roche UPL protocol (denature 94°C for 10 mins, cycle 94°C for 15 sec/60°C for 30 sec/72°C for 15sec (x45), cooling 40°C for 2 mins).

Results were analysed using SDS 2.4 (Thermo Fisher), Microsoft Excel (Microsoft Corporation, Washington, USA) and qbase+ software (Biogazelle, Gent, Belgium) (Hellemans *et al.*, 2007). Further details are provided for specific experiments in Table 2.4. Statistical analyses were performed using qbase+ and Excel. For the Chapter 3 qPCR panel (Figure 3.17G), paired *t* tests in qbase+ were used to calculate *p*-values, comparing -dox and +dox samples (3-4 pairs per cell line group). Correction for multiple comparisons was performed using the Benjamini-Hochberg method (Benjamini and Hochberg, 1995) and Microsoft Excel. For the Chapter 5 qPCR panel (Figure 5.18), statistical analysis was performed using qbase+ one-way ANOVA with correction for multiple hypotheses.

Additional information is provided in Table 2.4 (experimental details), and Tables 2.5-2.7 (assay details).

**Table 2.4: Quantitative PCR experiments**

Description	Associated figure/s	cDNA	Assays	Analysis	Other notes
Breast cancer cell lines, ELF5 total and ELF5 isoforms	3.12A	2ug of RNA in 20uL reaction volume. Diluted 1:5 in nuclease-free water.	Taqman assays ELF5 (total): Hs01063022_m1 ELF5 (isoforms 2 and 3 only): Hs00154971_m1 ELF5 (isoforms 1 and 4): custom assay GAPDH: 4326317E	The Pfaffl method (Pfaffl, 2001) was used by qbase+ to calculate relative quantities normalised to a single reference gene (GAPDH).	A custom Taqman assay was designed to detect Isoforms 1/4, using primers spanning the exon 2/3 boundary: GCCAGCTCTGAGAAGGGTTCA (forward primer), TGTGTGTCACCGAGTCCAACAT (reverse primer), and CTGTGGGAGTGAGGCAG (probe).
ELF5 isoforms cell lines +/- doxycycline	3.13G	0.5ug RNA in 20uL reaction volume. Diluted 1:5 in nuclease-free water.	Taqman assays ELF5 (total): Hs01063022_m1 GAPDH: 4326317E	Results analysed using SDS.24 software/Microsoft Excel and normalised relative quantities (calculated from Ct value using gene-specific standard curve).	
ELF5 isoforms cell lines qPCR panel	3.17	2.5ug RNA in 100uL reaction volume. Diluted 1:10 in nuclease-free water.	See Table 2.5 (Roche assays) and Table 2.6 (Taqman assays)	The Pfaffl method (Pfaffl, 2001) was used by qbase+ to calculate Normalized Relative Quantities (NRQ), which were normalized to a single reference gene (GAPDH) with error propagation. qPCR plates were laid out so that all samples for a single assay (in each qPCR round) were run on the same plate, known as a sample maximization approach (Hellemans <i>et al.</i> , 2007). To compare the results of assays run in both rounds 1 and 2 (on different plates), inter-run calibration was performed using qbase+ software, based on at least 3 identical samples that were run on both plates. This process calculates a calibration factor for each assay that corrects for any run-to-run differences, generating Calibrated Normalized Relative Quantity (CNRQ) values (Hellemans <i>et al.</i> , 2007).	48 hours of doxycycline treatment. Workflow shown in Figure 3.16.
DNA-PKcs knockdown optimisation experiment 1, MCF7-ELF5-V5 cells	5.12B	1ug RNA in 20uL reaction volume. Diluted 1:5 in nuclease-free water.	Taqman assays ELF5 (total): Hs01063022_m1 DNA-PKcs (PRKDC): Hs00179161_m1 GAPDH: 4352934	Results analysed using SDS.24 software/Microsoft Excel and normalised relative quantities.	

MCF7-, T47D- and MDA-MB-231- ELF5-V5 cell lines +/- doxycycline treatment and DNA-PKcs knockdown	5.17	<p>MCF7 lines: 3ug RNA in 60uL reaction volume (lot 1) or 1.5ug in 30uL reaction volume (lot 2).</p> <p>T47D lines: 3ug RNA in 60uL reaction volume.</p> <p>MDA-MB-231 lines: 1.5ug in 60uL reaction volume.</p> <p>All cDNA diluted 1:5 in in nuclease-free water.</p>	See Table 2.7	<p>The Pfaffl method (Pfaffl, 2001) was used by qbase+ to calculate Normalized Relative Quantities (NRQ), which were normalized to a single reference gene (GAPDH) with error propagation. qPCR plates were run in cell line groups, with all experimental and technical replicates for a single assay run on the same plate (sample maximisation approach) (Hellemans <i>et al.</i>, 2007).</p>	<p>Genes were selected based on a range of criteria (see Table 2.7), including significant expression changes in previous MCF7-ELF5-Isoform2-V5 Affymetrix arrays, significant expression changes in MCF7-ELF5-Isoform2-V5 RNA-sequencing, and the presence of an ELF5 ChIP-seq peak in the promoter region in MCF7-ELF5-Isoform2-V5 cells.</p>
--	------	---	---------------	--	---

**Table 2.5: Roche qPCR assays (related to Figure 3.17)**

Gene Symbol	Forward primer	Reverse primer	UPL probe	Amplicon (bp)	Intron-spanning (Y/N)	All isoforms (Y/N)	Test standard curve slope (efficiency)	Standard curve slope round 1 (efficiency)	Standard curve slope round 2 (efficiency)
ADAM17	cctttctgcgagagggaac	cacctgcaggaggtgtcagt	78	69	Yes	Yes	-3.3499 (98.85%)	-3.530 (91.99%)	
<b>AKT1</b>	ggctattgtgaaggagggtg	tcctgtagccaatgaagggtg	69	108	Yes	Yes	-3.3267 (99.80%)	-3.301 (100.88%)	
AREG	tgatcctcacagctgttct	tccattctctgtcgaagttct	73	107	Yes	Yes	-3.4377 (95.39%)	-3.481 (93.76%)	-3.259 (102.69%)
AURKA	gcagattttgggtggctcagt	tagtcagggtgccacaga	79	68	Yes	Yes	-3.5584 (91.00%)	-3.404 (96.69%)	
BCAS3	gtcaatcactcgggagact	gccatagcttcattcataaacc	78	93	Yes	Yes	-3.4419 (95.23%)	-3.851 (81.83%)	
BCL2L1 (BCLX)	gctgagttaccggcatcc	ttctgaaggagagaaaagagattc	10	124	Yes	Yes	-3.4027 (96.74%)	-3.49 (93.43%)	
BMP7	accactgggtggtaatcc	caactgggggtgatgtctct	42	86	Yes	Yes	-3.8826 (80.95%)	-3.381 (97.59%)	
BTN3A1	caagtttctggccttcctt	agagggtccaagcacagaaa	16	101	Yes	Yes	-3.2097 (104.91%)	-3.312 (100.42%)	
C3orf57	gtgaaggagcaaggctgaag	tcaagctgcagtaagttgtcc	85	63	Yes	Yes	-3.3369 (99.38%)	-3.369 (98.07%)	
C6orf192	tgttttagcaggaatgtttgc	ggaactcggccaatacacc	63	67	Yes	Yes	-3.2313 (103.93%)	-3.092 (110.58%)	
<b>CALCOCO1</b>	cagagtgggggtgaggag	gacagggtaccactgtaaagc	38	92	Yes	Yes	-3.3135 (100.35%)	-3.199 (105.40%)	-3.299 (100.97%)
CAMK2N1	tttatagggtctcttaaggcaca	gcattttgcaaataccatgc	27	61	No	Yes	-3.4003 (96.83%)	-3.529 (92.03%)	
CARM1	aaccacaccgactcaagga	aaaaacgacaggatccaga	76	68	Yes	Yes	-3.2189 (104.49%)	-3.151 (107.66%)	
CASP2	cgccatctatggtgtggat	ttctgtaggctgggcagtt	78	87	Yes (short intron-spanning)	Yes	-3.9051 (80.33%)	-3.569 (90.63%)	
CASP3	ttgtgaattgatgcgtgat	ggctcagaagcacacaaaca	68	75	Yes	Yes	-3.5702 (90.59%)	-3.347 (98.96%)	
<b>CASP4</b>	ttcctggcaattgaaaatgg	tgcaagctgtactaatgaagggtg	23	85	Yes	Yes	-3.0174 (114.49%)	-3.206 (105.08%)	-3.272 (102.13%)
CCNB1	acatggtgcacttcctcct	aggtaatgttagagttggtgtcc	18	103	Yes	Yes	-3.3251 (99.87%)	-2.899 (121.28%)	
CCNB2	tggaaaagttggctccaaag	tcagaaaaagctggcagaga	7	85	Yes	Yes	-3.1855 (106.03%)	-3.281 (101.74%)	

CCND1	gctgtgcatctacaccgaca	ttgagctgttcaccaggag	17	78	Yes	Yes	-3.555 (91.11%)	-3.714 (85.89%)	
CCNE2	gccattgattcattagattcca	ctgtccactccaaacctg	74	107	Yes	Yes	-3.2144 (104.69%)	-3.098 (110.28%)	
CCNG2	gggggtgtttgatgaaagt	gatcatggggaggagagctg	55	91	Yes	Yes	-3.2285 (104.05%)	-3.038 (113.39%)	
CDC25A	cgatcatgagaactacaaacctga	tctggctcttcaactgacc	67	96	Yes	Yes	-3.2714 (102.15%)	-3.191 (105.77%)	
CDK2	aaagccagaacaagtgacg	gtactggggcacacctcagt	77	83	Yes	2 of 3 transcripts	-3.7931 (83.50%)	-3.343 (99.13%)	
CHEK1	caacaacccctcaagaaagg	tggattgaatgtgcttagaaaatc	14	95	Yes	Yes	-3.4247 (95.88%)	-3.373 (97.91%)	
<b>CTGF</b>	cctgcaggctagagaagcag	tggagatttgggagtagcg	85	111	Yes	Yes	-3.2736 (102.06%)	-3.422 (95.99%)	-3.342 (99.17%)
CWC27	ggacacaagtcgaatgttaa	tgcttctttctgagcttca	78	88	Yes	Yes	-3.8076 (83.08%)	-3.611 (89.20%)	
<b>DDIT3</b>	aaggcactgagcgtatcatgt	tgaagatacacttcttgaaca	21	105	Yes	Yes	-3.441 (95.26%)	-3.392 (97.16%)	-3.861 (81.55%)
E2F1	tccaagaaccacatccagt	ctgggtcaaccctcaag	5	75	Yes	Yes	-3.2553 (102.86%)	-2.983 (116.39%)	
EHHADH	cagcttctcccagactcac	ttcatctgctaaaatacgtcttct	18	78	Yes	Yes	-3.1301 (108.69%)	-3.292 (101.26%)	
EPHA2	ccaggcaggctacgagaa	ggctctcagatgcctcaaac	88	80	Yes	Yes	-3.1544 (107.50%)	-3.206 (105.08%)	
<b>GAPDH*</b>	agccacatcgctcagacac	gcccatacagaccaaacc	60	66	Yes	3 of 4 transcripts	-3.2904 (101.33%)	-3.299 (100.98%)	-3.507 (92.82%)
<b>GDF15</b>	ccggatactcacgccaga	agagatacgcaggtgcaggt	28	61	Yes	Yes	-3.7778 (83.95%)	-3.825 (82.57%)	-4.172 (73.66%)
GPR81	ttgtcatgtgactgtgaattgaa	tgacctccactaaaaattacaca	87	105	No (has no introns)	Yes	-3.5658 (90.74%)	-3.541 (91.60%)	
GREB1	acaatggccacaatgctctt	tgattggagaattccgtgaag	76	91	Yes	Yes	-3.2824 (101.68%)	-3.398 (96.92%)	
HEY1	catacggcaggagggaag	gcactagtccttcaatgatgct	29	125	Yes	2 of 3 transcripts	-3.0756 (111.42%)	-2.898 (121.34%)	
HTRA2	agtcagtacaacttcatcgaga	ccgttcgagataggacctc	22	113	Yes	Yes	-3.6591 (87.63%)	-3.882 (80.97%)	
IGSF9	cctgggtgcatctgacagt	ccagtctctgacttccaa	18	78	Yes	Yes	-3.1551 (107.47%)	-3.247 (103.22%)	
<b>INSR</b>	gctggaattatgcctcaaagg	tgagaatctcagactcgaatgg	54	75	Yes	Yes	-3.4484 (94.98%)	-3.400 (96.84%)	-3.519 (92.39%)
IRAK1	tgctggtgtacggcttc	ctgaggccaggagagaggt	86	90	Yes	Yes	-3.607 (89.34%)	-3.757 (84.57%)	

IRAK4	tgatatttacagcttgggtg	ggttcacgggttcatcca	11	75	Yes	Yes	-3.3976 (96.94%)	-3.042 (113.17%)	
ITGB1	cgatgccatcatgcaagt	acaccagcagccgtgaac	65	71	Yes	Yes	-3.3523 (98.75%)	-3.332 (99.58%)	
<b>JMY</b>	gaaagactgctgaaggttgc	gaggctagcacctcatcat	63	64	Yes	Yes	-3.4679 (94.25%)	-2.936 (119.08%)	-3.083 (111.04%)
KDM4A	gccgctagaagttcagtgag	gcgtcccttgacttcttatt	53	90	Yes	Yes	-3.4963 (93.20%)	-3.251 (103.05%)	
LRRFIP1	tcatgtaccaggttgataccctaa	cgtgttttcccttcaaatc	67	114	Yes	Yes	-3.4989 (93.11%)	-3.386 (97.39%)	
LYN	aagttggtgaaaaggcttg	gccaccttggtactgtgtatag	60	77	Yes	Yes	-3.2247 (104.22%)	-3.014 (114.68%)	
MATN3	ttccaggaaacctctgtgc	tgtatccttggtacactcacagt	60	115	Yes	Yes	-3.778 (83.95%)	-3.419 (96.10%)	
MCM3	cgacgttattctgatctacca	caaggggattgtctctca	49	85	Yes	Yes	-3.2817 (101.71%)	-3.254 (102.91%)	
MCM4	tgttgctcacaatgatctcg	cgaataggcacagctcgata	64	85	Yes	Yes	-3.7822 (83.82%)	-3.684 (86.83%)	
MUM1L1	tcagagagttctcgaagattgg	gcacttctgtggtgacctg	45	66	Yes	Yes	-2.9797 (116.57%)	-3.217 (104.57%)	
PDPK1	cgaggactgctatggcaatta	ggaggctgacaggagtg	63	94	Yes	Yes	-3.3211 (100.03%)	-3.075 (111.45%)	
RAD51	tgagggtaccttaggccaga	cactgccagagagaccatacc	66	65	Yes	Yes	-3.0895 (110.71%)	-3.351 (98.80%)	
RB1CC1	catcttgagaatcaaatagcaaaaag	tttctgaagttcagcaactaagc	82	85	Yes	Yes	-3.638 (88.31%)	-3.23 (103.98%)	
RET	ctccgtggatgcctcaa	ccaagttctccaggggaat	65	62	Yes	Yes	-3.5321 (91.92%)	-3.469 (94.21%)	
RRAGD	ggctagcggactacggaga	ggggctactgaagtcagaaac	82	89	Yes	Yes	-3.5585 (90.99%)	-3.681 (86.92%)	
SIX1	gaaccggaggcaaagagac	ggagagagttggtctgctg	63	96	Yes	Yes	-3.4058 (96.62%)	-3.031 (113.76%)	
SLC6A14	tgaaatgcccagatttctc	ttctctgatgaagccgacact	7	61	Yes	Yes	-3.4219 (95.99%)	-3.262 (102.56%)	
<b>SNAI1</b>	gctgcaggactctaaccaga	atctccggagggtggatg	11	84	Yes	Yes	-3.3866 (97.37%)	-3.482 (93.73%)	-3.912 (80.15%)
<b>SNAI2</b>	tggttgcttaaggacacat	gcaaagtctctgttcagtg	7	77	Yes	Yes	-3.2942 (101.17%)	-3.348 (98.92%)	-3.998 (77.88%)
TFAP2A	acatgctctggtctacaaaac	aggggagatcggtctctga	62	70	Yes	Yes	-3.2704 (102.20%)	-3.307 (100.63%)	
TFAP2C	cgaagaggactgagga	ggggctgtagaggtgctg	32	94	Yes	Yes	-3.356 (98.06%)	-3.431 (95.64%)	
TFE3	gaggcaccacaggactgc	ttgactactgtacacatcaagcaga	27	74	Yes	Yes	-3.3306 (99.64%)	-3.777 (83.98%)	
<b>TGFB3</b>	aagaagcgggcttggac	cgcacacagcagttctcc	38	62	Yes	Yes	-3.0614 (112.15%)	-3.222 (104.35%)	-3.947 (79.21%)
<b>TLR2</b>	cgttctctcaggtgactgctc	ccttggatcctgcttgc	14	63	Yes	Yes	-3.7295 (85.41%)	-3.741 (85.06%)	-4.077 (75.91%)



TLR5	gacacaatctcggctgactg	tcaggaacatgaacatcaatctg	16	105	Yes	Yes	-3.054 (112.54%)	-3.324 (99.91%)	
TLR6	ctgccaagattcaggagtg	ccattgccttacaacaaagttct	52	63	No	Yes	-3.4706 (94.15%)	-3.252 (103.00%)	
<b>TMEM44</b>	cttcggcctgtggatctg	tgggtttctgtgcacatctc	18	83	Yes	Yes	-3.4495 (94.94%)	-3.566 (90.73%)	-4.598 (65.00%)
TOLLIP	aacctcgtcatgtcctacgc	gctggtacactgttggcatc	9	85	Yes	Yes	-3.143 (108.05%)	-2.929 (119.49%)	
TRAF6	tttggtgccatgaaaaga	ctcatgtgtgactgggtgttc	6	75	Yes	Yes	-3.1545 (107.49%)	-3.053 (112.59%)	
UBR4	caggaaccctctctgacacc	aaccatctgtcgtctgtga	19	75	Yes	Yes	-3.1329 (108.54%)	-3.198 (105.44%)	
ULBP3	aggaagaagaggctggaacc	ctatggcttgggttgagcta	57	70	Yes	Yes	-3.2897 (101.36%)	-3.4 (96.84%)	
WEE1	tctgcgtgggcagaagat	tctgtagtttctactatagcatcagc	31	90	Yes	Yes	-3.5242 (92.20%)	-3.47 (94.17%)	
WIPI2	tcaaactcgagactgtgaaagaa	agcactttcccgaagtaccc	77	73	Yes	Yes	-3.3368 (99.38%)	-3.315 (100.29%)	
ZEB1	ttttcctgaggcacctgaa	aaaatgcacatcgtgttccat	34	87	Yes	Yes	-3.0125 (114.76%)	-2.733 (132.22%)	
ZEB2	acaagccaggacagatca	gccacactctgtgcatttga	68	77	Yes	Yes	-3.352 (98.76%)	-3.531 (91.96%)	
ZMIZ1	caatgctgaaggggtgt	ctccagctccatcctgctt	55	80	Yes	Yes	-3.3102 (100.49%)	-3.095 (110.43%)	

**Table 2.5.** Details of the 75 Roche qPCR assays used in Figure 3.17 (part of the 116-gene panel). Gene names in bold indicate that the assay was included in the second qPCR round. Asterisk (\*) indicates that slope and efficiency for GAPDH assay in qPCR rounds 1 and 2 is the average of all plates in the round.

**Table 2.6: Taqman qPCR assays (related to Figure 3.17)**

Gene Symbol	Assay Number	Standard curve slope round 1 (efficiency)	Standard curve slope round 2 (efficiency)
AURKB	Hs00945858_g1	-3.505 (92.89%)	
<b>BBC3 (PUMA)</b>	Hs00248075_m1	-3.638 (88.31%)	-3.698 (86.39%)
BCL2	Hs00608023_m1	-3.58 (90.25%)	
BCL2L11 (BIM)	Hs00708019_s1	-3.783 (83.80%)	
C5orf41	Hs01078210_m1	-3.621 (88.87%)	
CASP8	Hs01018151_m1	-3.525 (92.17%)	
<b>CASP9</b>	Hs00609647_m1	-3.551 (91.25%)	-3.829 (82.46%)
CCDC88C	Hs00325884_m1	-3.485 (93.62%)	
CCNA2	Hs00153138_m1	-3.581 (90.22%)	
CDC2 (CDK1)	Hs00938777_m1	-3.404 (96.69%)	
CDH1	Hs01023894_m1	-3.537 (91.75%)	
CDK6	Hs01026371_m1	-3.488 (93.51%)	
CDKN1A (p21)	Hs00355782_m1	-3.555 (91.11%)	
CDKN2C (p18)	Hs00176227_m1	-3.518 (92.42%)	
CDKN3	Hs00937839_m1	-3.555 (91.11%)	
<b>DEDD2</b>	Hs00958058_m1	-3.869 (81.33%)	-3.866 (81.41%)
DKK1	Hs00183740_m1	-3.459 (94.58%)	
DUSP6	Hs04329643_s1	-3.676 (87.08%)	
<b>ELF5</b>	Hs01063022_m1	-3.508 (92.78%)	-3.599 (89.61%)
ESR1	Hs00174860_m1	-3.538 (91.71%)	
ETS1	Hs00428293_m1	-3.397 (96.96%)	
<b>FOXA1</b>	Hs00270129_m1	-3.536 (91.78%)	-3.476 (93.95%)
<b>GAPDH*</b>	4326317E	-3.426 (95.83%)	-3.425 (95.87%)
<b>GATA3</b>	H200231122_m1	-3.607 (89.34%)	-3.501 (93.03%)
HIP1	Hs00193477_m1	-3.379 (97.67%)	
<b>ID2</b>	Hs04187239_m1	-3.443 (95.18%)	-3.597 (89.67%)
ID4	Hs02912975_g1	-3.527 (92.10%)	
LUM	Hs00929860_m1	-3.955 (79.00%)	
MCL1	Hs01050896_m1	-3.649 (87.95%)	
<b>MYC</b>	Hs00906030_m1	-3.575 (90.42%)	-3.603 (89.47%)
NFKBIA	Hs00153283_m1	-3.483 (93.69%)	
<b>NUPR1</b>	Hs01044304_g1	-3.395 (97.04%)	-3.46 (94.54%)
PLK1	Hs00153444_m1	-3.684 (86.83%)	
<b>PMAIP1 (NOXA)</b>	Hs00560402_m1	-3.42 (96.06%)	-3.56 (90.94%)
RUNX2	Hs00231692_m1	-3.584 (90.12%)	
<b>SOCS1</b>	Hs00864158_g1	-2.957 (117.86%)	-3.193 (105.68%)
<b>SOCS2</b>	Hs00919620_m1	-3.462 (94.47%)	-3.811 (82.98%)
SOCS3	Hs01000485_g1	-3.835 (82.29%)	
STAT1	Hs01013996_m1	-3.563 (90.84%)	
<b>TLR3</b>	Hs01551078_m1	-3.607 (89.34%)	-3.908 (80.25%)

TP53	Hs01034249_m1	-3.552 (91.22%)	
<b>ULBP1</b>	Hs00360941_m1	-3.668 (87.34%)	-3.825 (82.57%)
VIM	Hs00185584_m1	-3.536 (91.78%)	
XRCC5	Hs00897854_m1	-3.509 (92.74%)	

**Table 2.6.** Details of the 44 Taqman qPCR assays used in Figure 3.17 (part of the 116-gene panel). Gene names in bold indicate that the assay was included in the second qPCR round. Asterisk (\*) indicates that slope and efficiency for GAPDH assay in qPCR rounds 1 and 2 is the average of all plates in the round.

**Table 2.7 (next page).** List of genes and Taqman assays selected for combined ELF5 overexpression and DNA-PKcs knockdown experiments. Genes were selected based on a range of criteria, including significant expression changes in previous MCF7-ELF5-Isoform2-V5 Affymetrix arrays (columns 4-5), significant expression changes in MCF7-ELF5-Isoform2-V5 RNA-sequencing (columns 6-7) and the presence of an ELF5 ChIP-seq peak in the promoter region in MCF7-ELF5-Isoform2-V5 cells (column 8). Significant (FDR<0.05) ELF5-induced expression changes are indicated by bold font (red = upregulation, green = downregulation). Expression changes with FDR 0.05-0.10 are indicated by non-bold red or green font. Information relating to quality control assessment of the assays are shown for each cell line where relevant in column 9. Affy, affymetrix microarray; FC, fold change; FDR, false discovery rate.

**Table 2.7: Quantitative PCR assays (related to Figure 5.18)**

Gene Symbol	Gene Name	Taqman Assay	Affy FC	Affy FDR	RNA-Seq FC	RNA-Seq FDR	Promoter Peak	Cell Line Notes
CDH1	Cadherin 1 (E-cadherin)	Hs01023894_m1	-1.103	0.093	-1.039	0.43	Yes	MDA-MB-231 lines excluded due to very low/absent expression in majority of samples
DKK1	Dickkopf WNT signaling pathway inhibitor 1	Hs00183740_m1	-2.231	4.7x10 <sup>-6</sup>	-2.62	7.1x10 <sup>-7</sup>	Yes	
ESR1	Estrogen receptor 1 (alpha)	Hs00174860_m1	-1.177	0.056	-1.005	0.98	No	MDA-MB-231 cells not measured as well-characterised as ER-negative
FILIP1L	Filamin A interacting protein 1 like	Hs00706279_s1	-1.292	0.013	-4.60	6.9x10 <sup>-5</sup>	Yes	
FOXA1	Forkhead box A1	Hs00270129_m1	-1.124	0.059	-1.036	0.58	Yes	
GATA3	GATA binding protein 3	Hs00231122_m1	-1.155	0.024	1.013	0.95	Yes	
GDF15	Growth differentiation factor 15	Hs00171132_m1	3.049	2.2x10 <sup>-5</sup>	2.791	3.0x10 <sup>-6</sup>	Yes	
GRHL3	Grainyhead like transcription factor 3	Hs00297962_m1	1.553	1.1x10 <sup>-4</sup>	2.291	0.0013	Yes	MDA-MB-231 lines excluded due to very low/absent and highly variable expression in majority of samples
LYN	LYN proto-oncogene, Src family tyrosine kinase	Hs00176719_m1	-2.013	3.4x10 <sup>-5</sup>	-2.117	5.8x10 <sup>-3</sup>	Yes	
MATN3	Matrilin 3	Hs00159081_m1	-1.660	3.3x10 <sup>-4</sup>	-3.942	6.0x10 <sup>-4</sup>	Yes	
PIP	Prolactin induced protein	Hs01114172_m1	1.103	0.45	9.753	4.4x10 <sup>-3</sup>	Yes	No expression detected in any of the MDA-MB-231 samples
SNAI2	Snail family transcriptional repressor 2	Hs00950344_m1	-3.465	8.5x10 <sup>-5</sup>	-4.548	3.1x10 <sup>-6</sup>	Yes	T47D lines excluded due to very low/absent and highly variable expression in majority of samples
SPDEF	SAM pointed domain containing ETS transcription factor	Hs00171942_m1	1.168	0.091	1.299	5.3x10 <sup>-4</sup>	Yes	MDA-MB-231 lines excluded due to highly variable expression
STAT1	Signal transducer and activator of transcription 1	Hs01013996_m1	-1.137	0.064	-1.415	3.0x10 <sup>-5</sup>	Yes	
VTCN1	V-set domain containing T cell activation inhibitor 1	Hs01552471_g1	4.080	5.2x10 <sup>-8</sup>	14.34	7.2x10 <sup>-5</sup>	Yes	MDA-MB-231 lines excluded due to very low/absent expression in majority of samples
ELF5	E74 like ETS transcription factor 5	Hs01063022_m1						
PRKDC (DNA-PKcs)	DNA-dependent protein kinase catalytic subunit	Hs00179161_m1						
GAPDH	Glyceraldehyde-3-phosphate dehydrogenase	Hs99999905_m1 or 4352934E						

### **RNA-sequencing (MCF7-ELF5-V5 cells)**

MCF7-ELF5-Isoform2-V5 cells ( $1 \times 10^6$ ) were seeded in T150 flasks, and doxycycline (or vehicle) treatment was commenced 24 hours after plating. Cell pellets were collected after 48 hours of doxycycline treatment and stored at  $-70^{\circ}\text{C}$ . RNA was extracted from thawed pellets with phenol/chloroform using the miRNeasy Mini Kit (Qiagen) with on-column DNase treatment. RNA quality was assessed on the 2100 Bioanalyser using RNA Nano chips (Agilent), with all samples having an RNA Integrity Number (RIN) of 9.8-9.9 (out of 10). RNA samples were submitted to The Ramaciotti Centre for Gene Function Analysis (UNSW, Sydney, Australia) for sequencing. Samples were prepared with the TruSeq Stranded Total RNA Sample prep kit (RS-122-2201 Illumina, San Diego, California, USA) according to the manufacturer's instructions. One  $\mu\text{g}$  of total RNA was used as input to the ribosomal ribozero RNA depletion, followed by 13 cycles of PCR to amplify the adapter-ligated cDNA. All 6 samples (three -Dox and three +Dox experimental replicates) were pooled in one lane. Sequencing was performed on the Illumina HiSeq2000 using v3 SBS reagents and 100bp paired-end reads. Demultiplexing of the samples was done with Casava 1.8.2 (Illumina).

### **DNA-Based Methods**

#### **ChIP-seq**

MCF7-ELF5-Isoform2-V5 cells ( $1.5 \times 10^6$ ) were seeded in 15cm plates, and doxycycline (or vehicle) treatment was commenced 24-48 hours after plating. After 48 hours of doxycycline treatment, cells were cross-linked for 10 minutes at room temperature using 1% formaldehyde diluted in cell growth medium. After 10 minutes, formaldehyde was quenched with 0.2M glycine. Plates were then placed on ice and washed twice with cold PBS. Cross-linked cells were collected in 2mL PBS using a cell scraper and pellets containing approximately 20 million cells were stored at  $-70^{\circ}\text{C}$ . Two biological replicates (each containing 1 x -Dox and 2 x +Dox pellets) were shipped to the laboratory of Dr Jason Carroll (Cancer Research UK Cambridge Institute), where FOXA1 and ELF5-V5 ChIP-seq were performed according to previously published protocols (Hurtado *et al.*, 2011; Kalyuga *et al.*, 2012).

## Computational Methods

### TCGA RNA-seq analysis

RNA-Seq Version 2 data for initial primary tumours and solid tissue normal samples (where  $n \geq 3$ ) were downloaded from TCGA data portal (<https://tcga-data.nci.nih.gov/tcga/>) (Brennan *et al.*, 2013; Cancer Genome Atlas Research Network, 2008, 2011, 2012a, 2012b, 2012c, 2013a, 2013b, 2014a, 2014b, 2014c, 2015a, 2015b; Cancer Genome Atlas Research Network *et al.*, 2013; Davis *et al.*, 2014), with institutional Human Research Ethics Committee exemption. Samples with available RNA-Seq Version 2 data (at August 2013 for breast and April 2014 for all other cancer types) were included. The RNA-Seq Version 2 TCGA pipeline for pre-processing of publicly available data used MapSplice (Wang *et al.*, 2010) for alignment and RSEM (Li and Dewey, 2011) for quantitation. Non-normalised gene and isoform data were downloaded from TCGA as RSEM expected (“raw”) counts, unadjusted for transcript length, and scaled estimates, adjusted for transcript length. Scaled estimates were multiplied by  $10^6$  to obtain transcripts per million (TPM) values. Normalised gene and isoform data were downloaded from TCGA as quantile normalised RSEM expected counts (unadjusted for transcript length), with the upper quartile set at 1000 for gene data and 300 for isoform data.

A summary of all TCGA samples used in the analysis is shown in Table 3.1. For breast cancer samples, PAM50 (Predication Analysis of Microarray 50-gene classifier) status was used to generate a subtyped cohort of 515 patients and 59 matched normal samples (Cancer Genome Atlas Research Network, 2012c; Parker *et al.*, 2009). Six additional normal samples, matching to tumours in the initial cohort, were included in differential expression analyses.

Limma voom (Law *et al.*, 2014) was used for differential expression analysis of gene-level RNA-sequencing data, with inputs as non-normalised gene data (RSEM expected counts). Filtering was applied to remove genes with low expression, keeping genes with count  $>1$  in at least  $n$  samples (where  $n$  = number of samples in smallest group of replicates). The Trimmed Mean of M-Values (TMM) normalisation method (Robinson and Oshlack, 2010) was applied followed by differential expression analysis using limma voom. All fold change (FC) and False Discovery Rate (FDR) values reported were generated by limma voom analyses. Clustered heat maps (Figures 3.10 and Additional Figure 3.2) were created using the R package ‘gplots’ (Warnes *et al.*, 2015).

As a comparison, differential expression analysis was also carried out using edgeR (McCarthy *et al.*, 2012; Robinson *et al.*, 2010; Robinson and Smyth, 2007, 2008; Zhou *et al.*, 2014a), with inputs as non-normalised gene data (RSEM expected counts) rounded to the nearest integer. Filtering was applied, keeping genes with count >1 in at least  $n$  samples (where  $n$  = number of samples in smallest group of replicates). A classic edgeR approach was used for analysis of unpaired data, while a glm approach was used for paired data.

Plots comparing TCGA breast cancer stage and *ELF5* expression (Figure 3.7) were generated using cBioPortal (Cerami *et al.*, 2012; Gao *et al.*, 2013).

### **Additional sequencing datasets analysis**

Additional RNA-seq datasets, including The Genotype-Tissue Expression (GTEx) dataset (GTEx Consortium, 2015) and the Illumina Human Body Map (Figure 3.5), were accessed through the EMBL-EBI Expression Atlas ([www.ebi.ac.uk/gxa/experiments](http://www.ebi.ac.uk/gxa/experiments)) using accession numbers E-MTAB-2919 and E-MTAB-513.

DNA-PKcs alterations (Figure 5.4G-H) were analysed using cBioPortal. Breast cancer molecular subtype information obtained from the METABRIC (Curtis *et al.*, 2012) and TCGA (Cancer Genome Atlas Research Network, 2012c) publications was used to generate a list of subtyped patient identifiers for input to cBioPortal.

### **MCF7-ELF5-V5 RNA-sequencing analysis**

Sequences were trimmed for adapters and quality using Fastq-Mcf. Alignment was done with STAR (v 2.4.0d) (Dobin *et al.*, 2013) against the human genome (hg38) with gencode v20 annotations. Transcript counts were summarised and transcripts per million (TPM) calculated using RSEM (v 1.2.18) (Li and Dewey, 2011). Counts were normalised using TMM (Robinson and Oshlack, 2010) and transformed using Voom (Law *et al.*, 2014). Differential expression analysis was carried out using limma (Smyth, 2004). Alignment and differential expression analysis were performed by Dr Daniel Roden (Garvan Institute of Medical Research).

Gene Set Enrichment Analysis (GSEA) (Subramanian *et al.*, 2005) was run in GenePattern (Reich *et al.*, 2006) in pre-ranked mode using a ranked list of the limma moderated t-statistics. One thousand gene-set permutations were performed using minimum and maximum gene-set sizes of 15 and 500, respectively. Gene-sets used in GSEA were extracted from version 6.0 of the Broad Institute's Molecular Signatures

Database (MSigDB) (Liberzon *et al.*, 2015) and extended with additional curated gene-sets from literature, previously used for ELF5 microarray analysis (Kalyuga *et al.*, 2012). Network-based visualization and analysis of the GSEA results was carried out using the Cytoscape (Shannon *et al.*, 2003) Enrichment Map (Merico *et al.*, 2010) plugin, with thresholds of: FDR (Q-value) = 0.05 (or 0.10 where indicated); p-value = 0.005; and overlap coefficient cut-off = 0.5. Gene set clusters were manually annotated with functional themes.

### **FOXA1 and ELF5 ChIP-seq analysis**

Single-end sequencing (36bp) was carried out and alignments were generated using the "Bowtie for Illumina" (v0.12.7) tool on Galaxy (Afgan *et al.*, 2016), mapped to the hg19 canonical female genome (Langmead *et al.*, 2009). Peaks were called using two peak callers: (i) MACS v1.4.1 with default parameters (Zhang *et al.*, 2008); and (ii) HOMER v4.0 with default parameters (Heinz *et al.*, 2010). Replicate and peak caller consensus peaks were identified using BedTools (Quinlan and Hall, 2010).

Transcription factor DNA binding motifs were identified using MEME-ChIP (Machanick and Bailey, 2011). Alignment and motif analysis were performed by Dr Daniel Roden (Garvan Institute of Medical Research).

ChIP-seq data was visualised using Integrative Genomics Viewer (Robinson *et al.*, 2011; Thorvaldsdóttir *et al.*, 2013). Chromatin states were analysed using ChromHMM (Ernst and Kellis, 2012) data from MCF7 cells (Taberlay *et al.*, 2014) and genomic binding sites were analysed using the *Cis*-regulatory Element Annotation System (CEAS) tool on Galaxy. "One-condition-only" peaks were defined as described in the main text (Figure 4.16B).

The transcription factor binding intensity heatmaps were generated by Dr Daniel Roden using deepTools (Ramirez *et al.*, 2014). First, normalized signal binding coverage was generated for each transcription factor using bamCoverage with `--extendReads` equal to the MACS fragment length, `--binSize 10` and `--normalizeTo1x 2451960000`. Then, the computeMatrix with `--missingDataAsZero` and plotHeatmap and plotProfile functions were used to generate the transcription factor heatmaps and signal binding profiles.

Functional analyses of ChIP-seq data were performed using the online tool Genomic Regions Enrichment of Annotations Tool (GREAT) (McLean *et al.*, 2010). All consensus ChIP-seq peaks were submitted for analysis. Gene regulatory domains



were assigned using the basal plus extension rule, assigning each gene a basal regulatory domain (default settings of 5.0kb upstream and 1.0kb downstream), which is extended in both directions to the nearest gene's basal domain but no more than the maximum extension (1000kb) in one direction. Each ChIP-seq peak was then associated with all genes whose regulatory domains overlapped with the region of binding. Due to the large size of the ELF5 and FOXA1 ChIP-seq datasets, statistical significance of enriched terms was assessed using the binomial test over genomic regions, due to the saturation of the hypergeometric test over genes that can occur with large datasets. For analysis of smaller ChIP-seq subsets, statistical significance was assessed using both tests.

Additional functional analyses of the ELF5 ChIP-seq data were performed using Enrichr (Chen *et al.*, 2013a; Kuleshov *et al.*, 2016). The top 4,000 peaks (ranked by MACS score) were submitted for analysis; this number was selected to remain within the automatic capping of 2,000 target genes that is applied by the Enrichr program.

### **Association of ChIP-seq peaks with direct target genes**

For the ELF5 ChIP-seq, a probable list of direct target genes was generated by: (i) Assigning each ChIP-seq peak to the closest gene; (ii) Filtering to include only peaks where the closest gene is within 10kb of the transcription start; (iii) Overlap of this gene list with differentially expressed genes identified in the ELF5 RNA-seq, defined by FDR <0.05 and absolute fold change >1.5. For the FOXA1 peaks gained or lost on ELF5 over-expression, lists of potential target genes were generated using GREAT (described above).

### **Functional analyses of gene and protein lists**

Differentially expressed gene lists were analysed for functional enrichments using MSigDB v6.0 GO Biological Process and Hallmark Collection gene sets (Liberzon *et al.*, 2015). RIME candidate proteins were analysed using the Database for Annotation, Visualization and Integrated Discovery (DAVID) v6.8 (Huang da *et al.*, 2009a, 2009b).

### **Gene ID conversions**

Gene identifiers (including Affy probe IDs, Ensembl gene IDs, and HGNC symbols) were converted using Ensembl Biomart (Kinsella *et al.*, 2011) and the DAVID gene ID conversion tool.

### **Correlation analysis**

Correlation analyses were performed in GraphPad (Prism) using Spearman rank-order or Pearson correlation as indicated.

### **Graphics**

Venn diagrams were created using online software (<http://bioinformatics.psb.ugent.be/webtools/Venn/>) or BioVenn (Hulsen *et al.*, 2008). Wordclouds for enriched pathways were generated online (<http://www.wordclouds.com>), with word size proportional to the number of occurrences (minimum 2).

### **Phosphosite prediction**

The ELF5 protein sequence was analysed for potential phosphorylation sites using Scansite 3 (<http://scansite3.mit.edu>) (Obenauer *et al.*, 2003). Predictions were made using high stringency (top 0.2% of motif matches within the vertebrate database), medium stringency (top 1%) and low stringency (top 5%) settings.

### **Survival analysis**

Km-plotter (Györfy *et al.*, 2010) was used to analyse the association between DNA-PKcs expression and breast cancer survival. Patients were divided by high and low expression of DNA-PKcs using a cut-off automatically generated by the Km-plotter program (involving assessment of all percentiles between the upper and lower quartiles and selection of the best performing threshold). Both overall survival (time from diagnosis to death from any cause) and disease-free survival (time to disease relapse) were analysed. Overall survival was also analysed in the METABRIC (Curtis *et al.*, 2012; Pereira *et al.*, 2016) and TCGA (Cancer Genome Atlas Research Network, 2012c) cohorts, using cBioPortal (Cerami *et al.*, 2012; Gao *et al.*, 2013). Patients were divided by alterations in DNA-PKcs mRNA expression (defined as z-score greater than 2.0 or less than -2.0 compared to the expression distribution for samples diploid for DNA-PKcs); the vast majority of altered expression was DNA-PKcs mRNA upregulation, and the small number of cases with mRNA down-regulation were excluded from the survival analyses. Individual subtype analysis was not performed for the TCGA cohort due to the small patient numbers.

## Chapter 3: ELF5 Isoforms

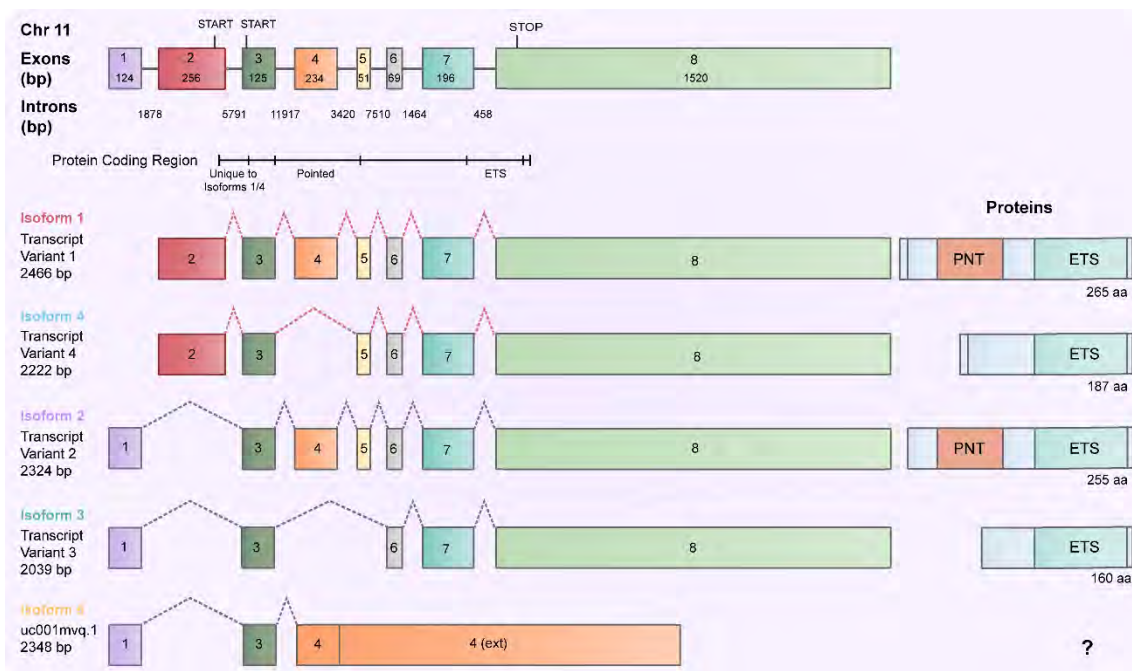
### Introduction

It is becoming increasingly recognised that almost all multi-exon genes undergo alternative transcription (such as alternative transcription start or termination sites) and/or alternative splicing, increasing diversity of protein structure and function (Pal *et al.*, 2012). Alternative transcription and splicing are also commonly deregulated in cancer, contributing to tumour initiation and progression but also providing potential cancer-specific therapeutic targets. Importantly, different protein isoforms produced by the same gene may have very different functions; one striking example is Vascular Endothelial Growth Factor (VEGF), which produces both pro-angiogenic and anti-angiogenic isoforms (Bates *et al.*, 2002). Early studies described tissue-specific differences in ELF5 transcript isoform expression (Oettgen *et al.*, 1999) but recent studies have not distinguished between isoforms or have used a single isoform for over-expression studies. This study (published Piggini *et al.*, 2016) represents the first comprehensive analysis of ELF5 expression at the isoform level, using RNA-sequencing data from The Cancer Genome Atlas (TCGA) for 6,757 normal tissue and cancer samples. The functional effects of ELF5 isoform expression in breast cancer were also investigated using inducible cell line models, leading to unique insights into the transcriptional functions of ELF5 and, in particular, the role of the Pointed domain.

## Results

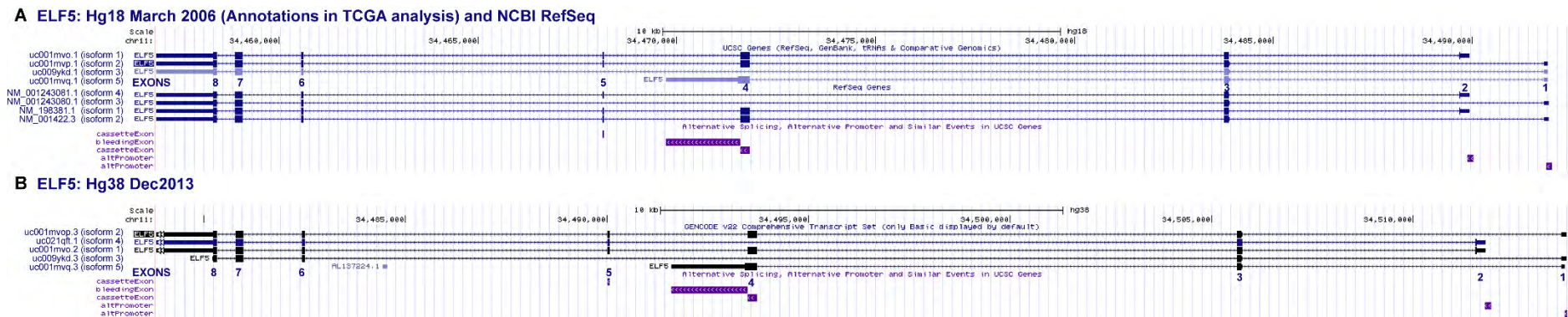
### **ELF5 isoforms are differentially expressed in normal tissues**

There are 4 ELF5 transcript variants in the NCBI RefSeq database (National Center for Biotechnology Information, 2002), predicted to produce 4 unique proteins (Figure 3.1). The two full-length transcripts (Isoforms 1 and 2) utilise alternative promoters, resulting in unique first exons and proteins that differ by only 10 N-terminal amino acids. Two additional transcripts (Isoforms 3 and 4) are produced by splicing of exons 4 (+/- 5) from each of the full-length transcripts, producing proteins that lack the Pointed domain but retain the ETS domain. An additional transcript (Isoform 5), described by Gencode (Harrow *et al.*, 2012), is a variant of Isoform 2 terminating at an extended exon 4. This type of intronic extension (“bleeding exon”) is often associated with incompletely processed transcripts (Kent *et al.*, 2002) and it is unclear whether this transcript produces a protein product (which would lack the DNA-binding ETS domain). Figure 3.1 also introduces the colour-coding that will be used for ELF5 isoforms throughout this thesis; ELF5 Isoform 2, for example, will be represented by purple in all chapters.



**Figure 3.1: ELF5 isoforms are produced by alternative promoter usage and splicing**

RefSeq and Gencode transcripts with protein products. ETS, E26 DNA-binding domain; PNT, Pointed domain; bp, base pairs; aa, amino acids; ext, extended.



**Figure 3.2: *ELF5* annotations**

UCSC genome browser screenshot with annotations showing (A) *ELF5* transcripts in Hg18 March 2006 and NCBI RefSeq and (B) the more recent transcripts in Hg38 December 2013. Hg18 transcript names match those that appear in The Cancer Genome Atlas RNA-sequencing analysis files. The more recent Hg38 includes an equivalent for NCBI *ELF5* Isoform 4 that does not appear in Hg18.

RNA-sequencing data from TCGA were analysed to quantify and compare *ELF5* isoforms in normal tissues and cancer (Brennan *et al.*, 2013; Cancer Genome Atlas Research Network, 2008, 2011, 2012a, 2012b, 2012c, 2013a, 2013b, 2014a, 2014b, 2014c, 2015a, 2015b; Cancer Genome Atlas Research Network *et al.*, 2013; Davis *et al.*, 2014). A summary of all TCGA samples analysed is shown in Table 3.1. TCGA pre-processed data include *ELF5* Isoforms 1, 2 and 3 as annotated by RefSeq, as well as Isoform 5. Due to the reference annotation used by TCGA there is no data for *ELF5* Isoform 4. The transcripts and protein products are summarised in Figure 3.1 and a cross-database comparison is shown in Figure 3.2.

*ELF5* expression was highest in epithelial tissues such as the breast, kidney, lung, prostate and bladder (Figure 3.3A). The breast was one of the highest *ELF5*-expressing tissues in the body. Isoform 1 and 2 expression was highly tissue-specific (Figure 3.3B), indicating alternative promoter use in different tissues.

Data in Figures 3.3A and 3.3B were quantile-normalised by The Cancer Genome Atlas pipeline, allowing comparison of abundance of a particular transcript (such as total *ELF5*) between samples. However, longer transcripts will generate more sequencing reads, making quantitative comparison of transcripts of different lengths problematic. To overcome this, the proportional measure 'transcripts per million' (TPM) may be used. TPM is an example of a within sample normalisation method and it should be noted that values are not technically comparable between samples, particularly when the composition of the total mRNA pool may be quite different (for example, when comparing different tissues). For this reason, both quantile normalised (between sample normalised, Figure 3.3B) and TPM normalised (within sample normalised, Figure 3.3C) data are shown. As the lengths of *ELF5* transcripts are not widely different, ranging from 2,039 to 2,466 base pairs, the data plots are in fact similar.

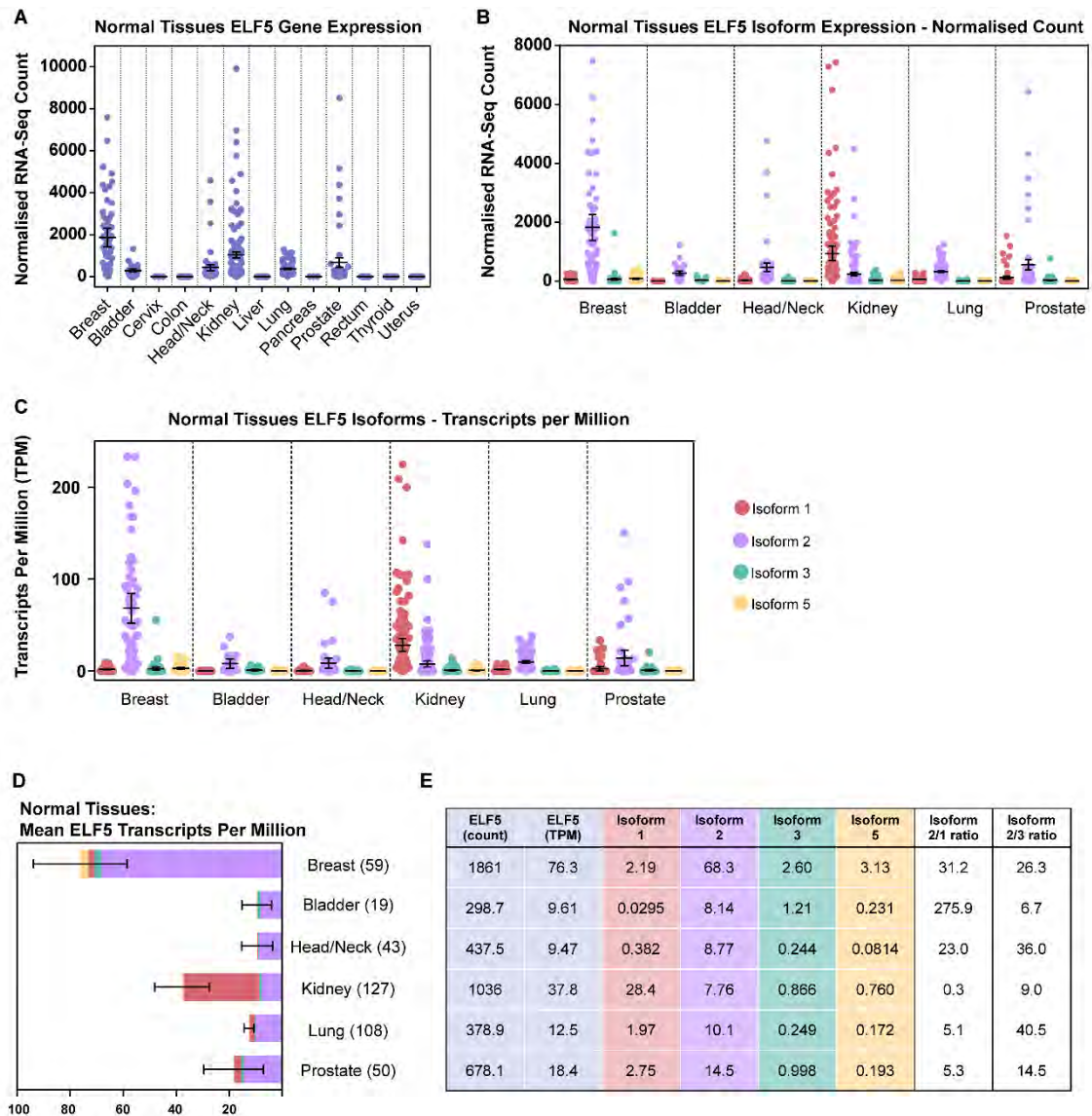
As TPM is a proportional measure, the relative abundances of transcripts of different lengths within samples can be compared. The mean TPM values for *ELF5* isoforms are shown in Figures 3.3D and 3.3E. Breast, bladder, head/neck, lung and prostate all expressed Isoform 2 as their main transcript (median percentage 82.1-95.2%), while the kidney expressed mainly Isoform 1 (median 91.8%). All tissues examined expressed on average more full-length Isoform 2 than the shorter Isoform 3. The isoform percentage values for each tissue are shown in Table 3.2.

**Table 3.1: Summary of all TCGA RNA-sequencing samples used in analysis**

Tissue	Cancer type	TCGA acronym	Normal samples <sup>a</sup>	Cancer samples
Bladder	Bladder urothelial carcinoma	BLCA	19	241
Breast	Breast invasive carcinoma	BRCA	59 <sup>b</sup>	515
	<i>Luminal A</i>			229
	<i>Luminal B</i>			126
	<i>HER2</i>			57
	<i>Basal-like</i>			96
	<i>Normal-like</i>			7
Cervix	Cervical squamous cell carcinoma and endocervical adenocarcinoma	CESC	3	185
Colon	Colon adenocarcinoma	COAD	41	261
Head/neck (including mouth and throat)	Head and neck squamous cell carcinoma	HNSC	43	497
Kidney	Chromophobe	KICH	25	66
	Clear cell carcinoma	KIRC	72	518
	Papillary cell carcinoma	KIRP	30	172
Liver	Hepatocellular carcinoma	LIHC	50	191
Lung	Lung adenocarcinoma	LUAD	58	488
	Lung squamous cell carcinoma	LUSC	50	490
Pancreas	Pancreatic adenocarcinoma	PAAD	3	85
Prostate	Prostate adenocarcinoma	PRAD	50	297
Rectum	Rectum adenocarcinoma	READ	9	91
Thyroid	Thyroid carcinoma	THCA	59	498
Uterus	Uterine corpus endometrial carcinoma	UCEC	24	158
	Uterine carcinosarcoma	UCS	NA <sup>c</sup>	57
Adrenal gland	Adrenocortical carcinoma	ACC	NA	79
Haematological	Diffuse large B-cell lymphoma	DLBC	NA	28
	Acute myeloid leukemia	LAML		173
Brain	Glioblastoma multiforme	GBM	NA	156
	Lower grade glioma	LGG		463
Ovary	Ovarian serous cystadenocarcinoma	OV	NA	262
Skin	Cutaneous melanoma	SKCM	NA	82
Bone/connective tissue/soft tissue	Sarcoma	SARC	NA	103

List of all TCGA samples used in ELF5 mRNA expression analysis. Breast cancer samples are sub-divided into molecular subtypes (according to Cancer Genome Atlas Research Network, 2012c), with the number of each subtype shown in italics (515 total).

<sup>a</sup> Normal samples included where  $n \geq 3$ . <sup>b</sup> 65 samples included in differential expression analysis. <sup>c</sup> UCEC normal samples used as normal uterine samples for differential expression analysis.



**Figure 3.3: *ELF5* isoforms are differentially expressed in normal tissues (quantile normalised counts)**

Plotted values represent individual TCGA RNA-sequencing samples and error bars the mean with 95% confidence interval. (A) *ELF5* gene expression in 13 normal tissues (quantile normalised counts). (B) *ELF5* isoform expression in selected normal tissues (quantile normalised counts). (C) Equivalent to graph shown in panel B using 'Transcripts per million' (TPM) values instead of quantile normalised RNA-Seq counts. Plotted values represent individual samples and error bars show the mean with 95% confidence interval. (D) Mean *ELF5* levels (Transcripts per Million, TPM) in normal tissues. Relative isoform contributions shown within each bar. Numbers in parentheses indicate samples per group. (E) Mean *ELF5* gene and isoform expression in normal tissues. All values are TPM, except for column 1 which is the quantile normalised count. Isoform ratios in final 2 columns calculated using mean TPM values.



**Table 3.2: *ELF5* splice variant proportions in normal tissues (based on TPM values)**

	Mean splice variant percentage				Median splice variant percentage			
	TV1	TV2	TV3	Other	TV1	TV2	TV3	Other
Bladder	0.2%	68.1%	24.0%	7.8%	0.0%	86.9%	11.5%	0.9%
Head/Neck	1.8%	90.6%	6.5%	1.0%	0.0%	95.2%	0.0%	0.8%
Kidney	86.3%	9.4%	2.3%	2.1%	91.8%	3.9%	0.9%	1.8%
Lung	14.1%	81.1%	3.3%	1.5%	13.9%	82.1%	0.0%	1.2%
Prostate	3.2	82.	12.	1.7	0.0	89.	1.4	1.3

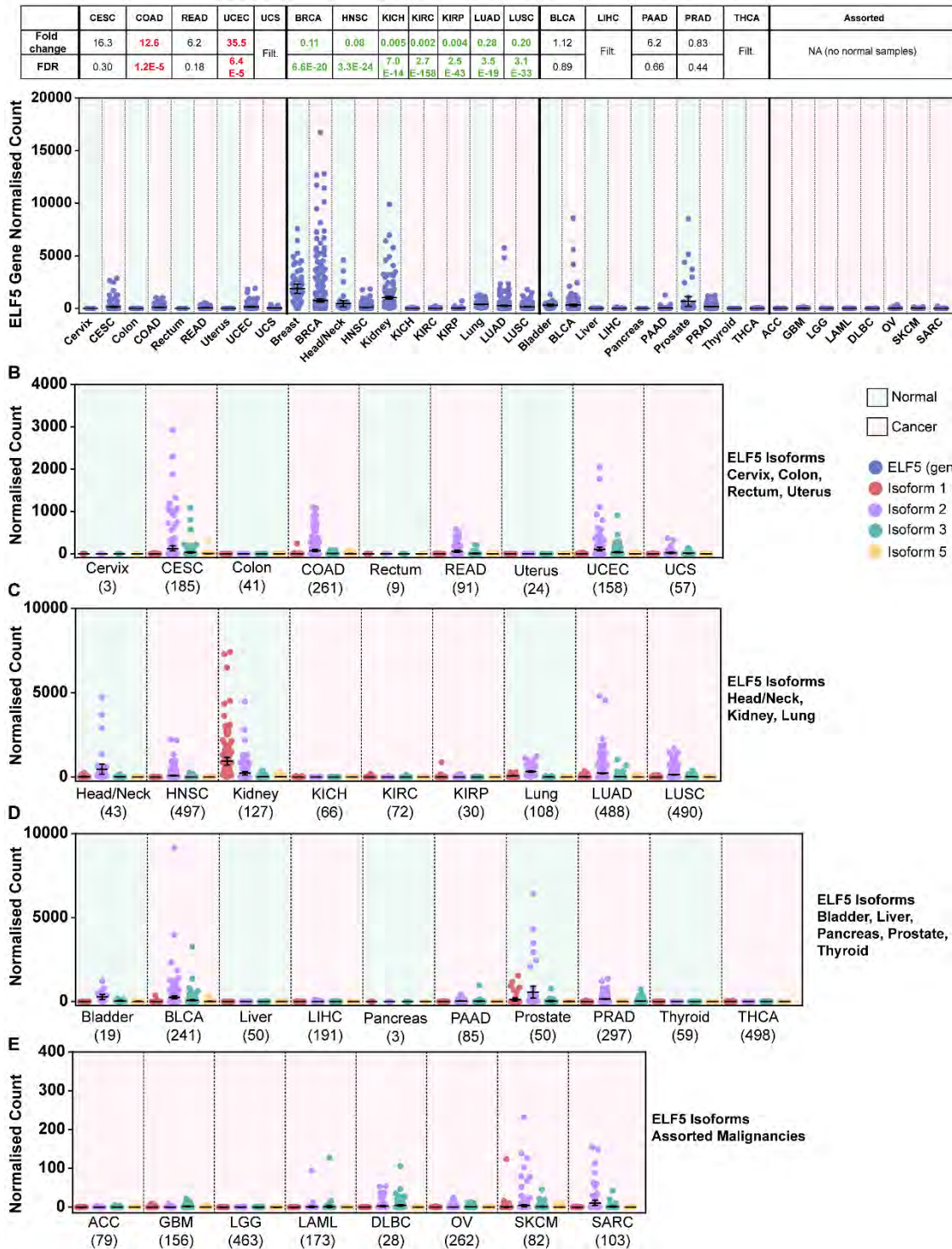
Mean (left) and median (right) percentage values for *ELF5* isoforms in selected normal tissues.

### ***ELF5* expression is significantly altered in cancer**

In malignancy, *ELF5* expression was significantly altered compared to normal, as shown by limma voom differential gene expression analysis (Figure 3.4A). In the cervix, colon, rectum and uterus, cancer was associated with an increase in *ELF5* level, driven mainly by an increase in Isoform 2 and to a lesser extent Isoform 3 (Figure 3.4B). Conversely, there was almost complete suppression of *ELF5* expression in 3 kidney carcinoma subtypes. *ELF5* expression was also significantly decreased in head/neck, lung and prostate cancer (Figure 3.4C). In both lung carcinoma subtypes, there was a large variation in *ELF5* level, suggesting possible molecular subtype-specific expression patterns, similar to the breast. *ELF5* expression was largely unchanged (or filtered from analysis due to low expression) in the tissues shown in Figure 3.4D. The cancer types shown in Figure 3.4E exhibited very low levels of *ELF5* expression but had no normal tissue samples available as a comparison. The TPM normalised (within sample normalised) values are shown in Figure 3.8A-C as a direct comparison to the quantile normalised (between sample normalised) values in Figures 3.4B-D.

Differential expression analysis was also carried out using edgeR. Overall, the results from limma voom and edgeR were very similar. The edgeR fold change and false discovery rate values are presented at the end of the chapter in Additional Figure 3.1A for comparison.

A





### ***ELF5* expression is altered in breast cancer in a subtype-specific manner**

Comprehensive analysis of RNA-sequencing incorporating molecular subtype was undertaken for 515 breast cancer patients. In the luminal A, luminal B and HER2 subtypes, *ELF5* was significantly downregulated (0.02-0.13 of normal), while in the basal subtype there was a strong trend for increased *ELF5* expression (1.96-fold compared to normal, FDR 0.053 in limma voom analysis, 1.99-fold compared to normal, FDR 0.0008 in edgeR analysis) (Figure 3.6A and Additional Figure 3.1B). The TPM normalised values are shown in Figure 3.8D.

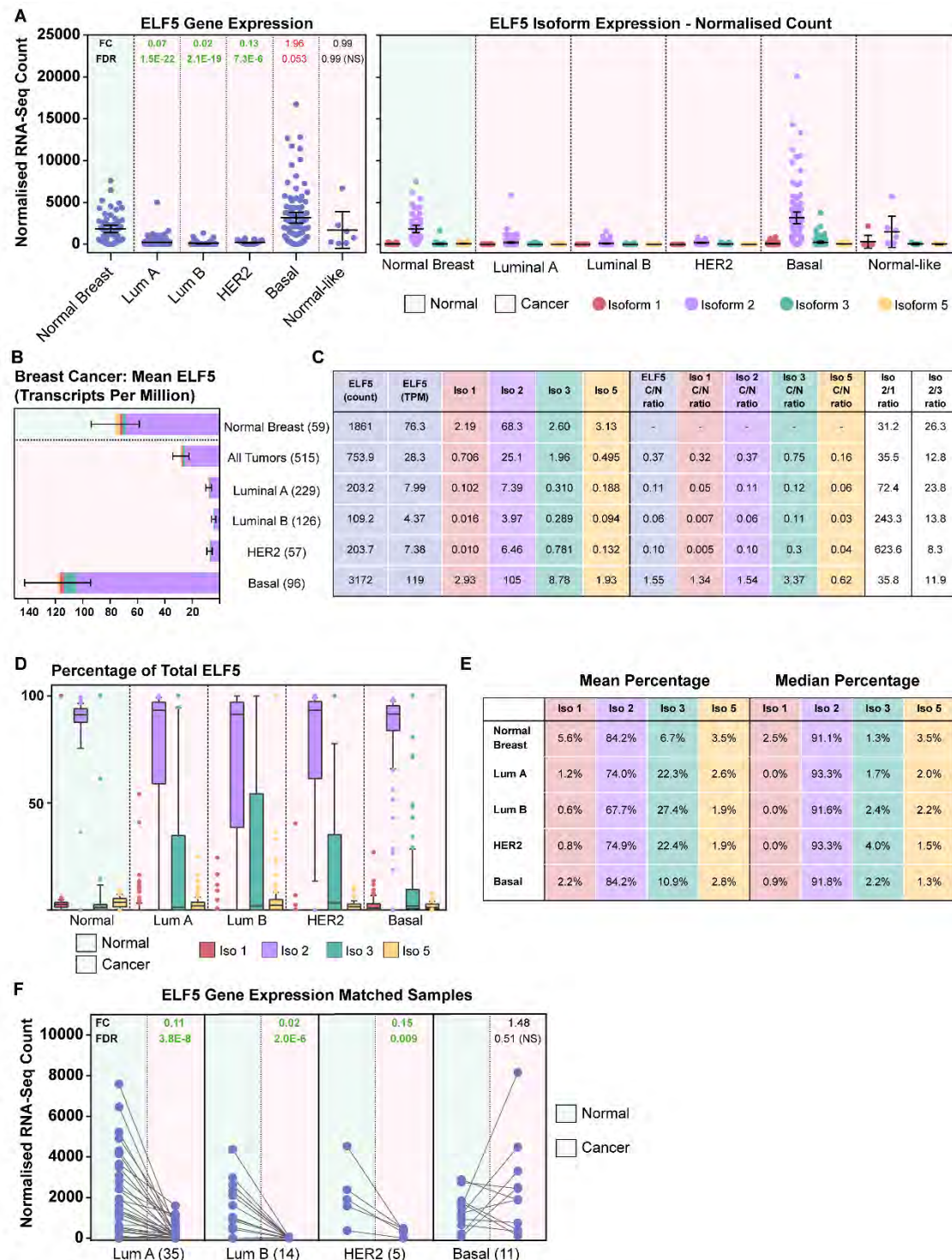
This analysis was extended to the isoform level by examining the contribution to total *ELF5* (based on mean TPM) for each isoform (Figure 3.6B). Normal-like samples were excluded due to low sample numbers. The main isoform expressed in all breast cancer subtypes was Isoform 2. In the luminal A, luminal B and HER2 subtypes, all *ELF5* isoforms were decreased in cancer compared to normal (Figure 3.6C). Conversely, in the basal subtype, 3 of 4 isoforms were upregulated, with Isoform 3 having a relatively larger fold change.

The percentage contributions of each isoform to total *ELF5* were also analysed (Figures 3.6D and 3.6E). The normal breast showed a tight range of expression, while in cancer, particularly for Isoforms 2 and 3, this was broadened (Figure 3.6D). The high variability in Isoform 3 percentage values in the cancer samples led to an increased mean percentage in all subtypes. Median values demonstrated a smaller, although still increased, Isoform 3 percentage in cancer, while the median Isoform 2 percentage remained fairly constant across normal and cancer samples.

Within this cohort, 65 patients had matched tumour and normal samples that could be directly compared (Figure 3.6F and Additional Figure 3.1C). The luminal A, luminal B and HER2 groups showed a highly significant decrease in *ELF5* level in both the limma and edgeR analyses. In the basal subgroup, there was an upward but variable trend.

There was no clear relationship between *ELF5* expression and American Joint Committee on Cancer (AJCC) stage (Figure 3.7).





**Figure 3.6: *ELF5* expression is altered in breast cancer in a subtype-specific manner**

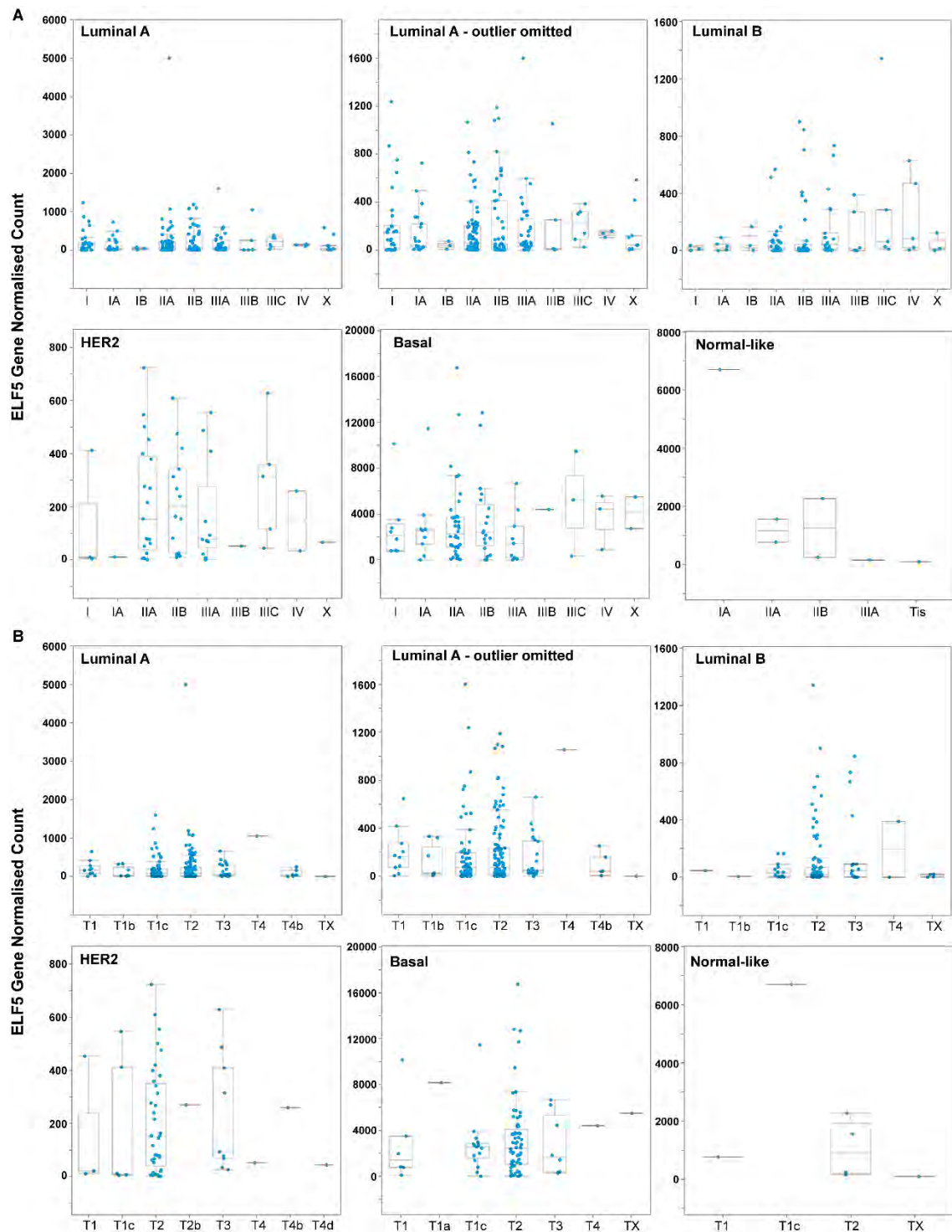
(A) *ELF5* gene (left) and isoform (right) expression (quantile normalised counts) for normal breast and breast cancer subtypes. Plotted values represent TCGA RNA-sequencing samples and error bars the mean with 95% confidence interval. Fold change (FC) and False Discovery rate (FDR) from limma voom analysis are shown for *ELF5* gene data, with green values in bold indicating a significant down-regulation and red values in bold a significant up-regulation compared to normal (FDR<0.05). Non-bold green or red values indicate FDR

0.05-0.10. (B) Mean *ELF5* levels (TPM) in normal breast and breast cancer, excluding normal-like, with 95% confidence interval. Relative isoform contributions shown within each bar. Numbers in parentheses indicate samples per group. (C) Mean *ELF5* expression values at the gene and isoform level (columns 1-6), isoform fold changes in cancer compared to normal (columns 7-11) and isoform ratios (columns 12-13). All values are TPM, except for column 1 which is the quantile normalised count. Ratios calculated using mean TPM values. (D) Box and whisker plot representing isoform percentage of total *ELF5* in normal breast and cancer. Box = 25-75th percentile, horizontal line = median, error bars = 10-90th percentile, circles = outliers. (E) Mean (left) and median (right) isoform percentage values for normal breast and cancer. (F) *ELF5* levels (quantile normalised count) for patients with matched normal and cancer samples, categorised according to tumour molecular subtype. Six extra matched normal samples were included for a total of 65 pairs. Plotted values represent individual samples, with samples from the same patient connected with a line. Fold change (FC) and False Discovery rate (FDR) from paired limma voom analysis are shown, with green values indicating a significant downregulation compared to normal (FDR<0.05). Numbers in parentheses indicate sample pairs per group.

### **Expression of other ETS family members is also altered in breast cancer, with the basal subtype having a distinct ETS expression profile**

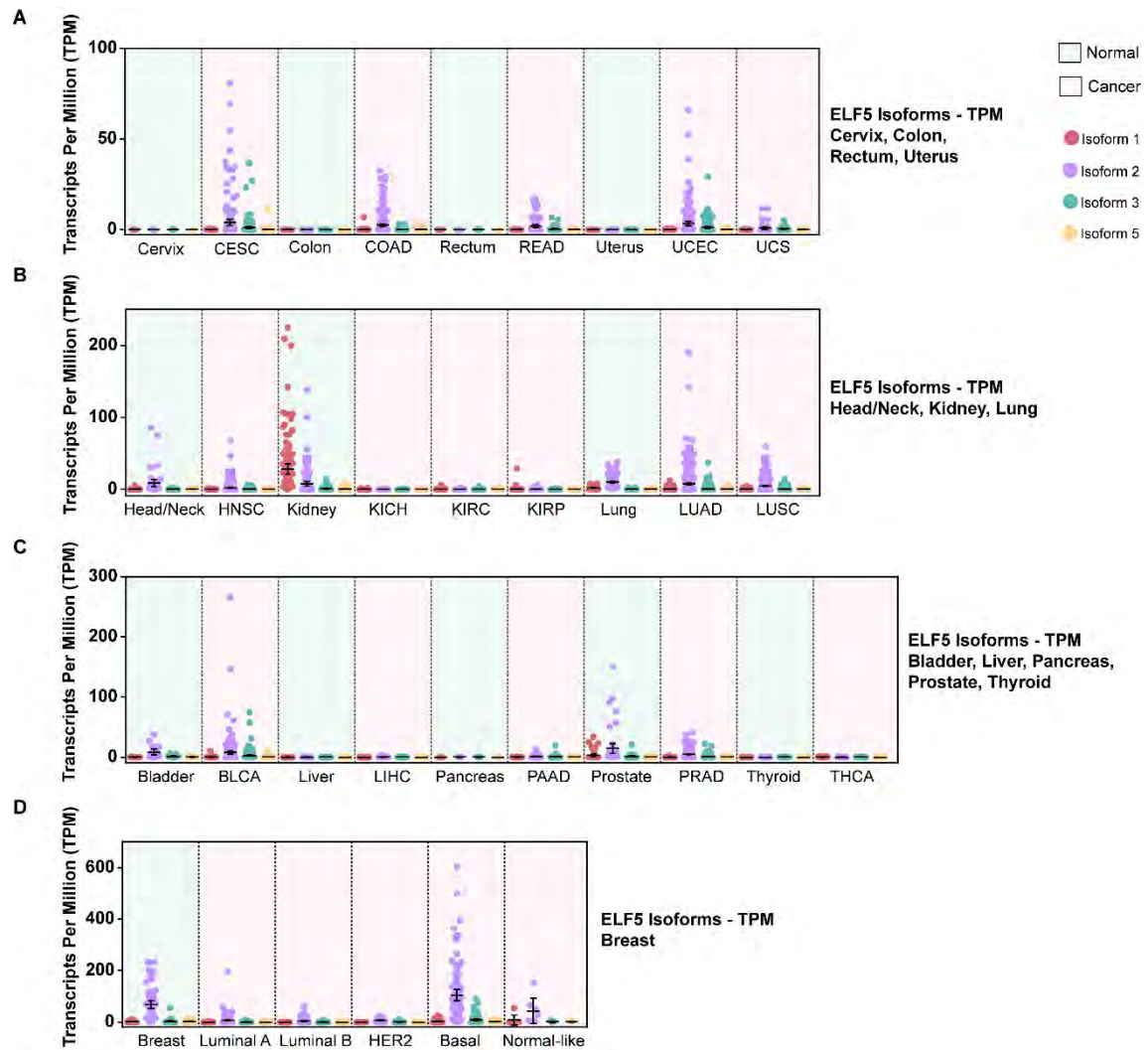
The same cohort of patients was used to examine expression of other members of the ETS transcription factor family. RNA-seq data showed that a large number of ETS factors were expressed in the normal breast. Average TPM values (which take into account transcript length) for ETS factors in the normal breast ranged from 0.02 to 117.7. Several ETS factors had very low expression (<2 TPM), including *FEV*, *SPIC*, *ETV2*, *ETV3L* and *SPIB*. The most highly expressed ETS factors in the normal breast were *EHF*, *ELF3*, *SPDEF* and *ELF5* (Figure 3.9, top).

ETS factor expression was significantly altered in breast cancer, as shown by limma voom differential expression analysis. In the first (unpaired) analysis, samples from each molecular subtype, excluding normal-like, were compared to the common set of 65 normal breast samples, allowing analysis of larger sample sets. In the second (paired) analysis, normal and subtyped tumour samples from the same patient were compared, allowing for more rigorous matched comparisons but limited by smaller sample numbers. ETS factors with low expression (3-5 per subtype) were filtered from the analysis.



**Figure 3.7: Stage compared to *ELF5* expression in breast cancer subtypes**

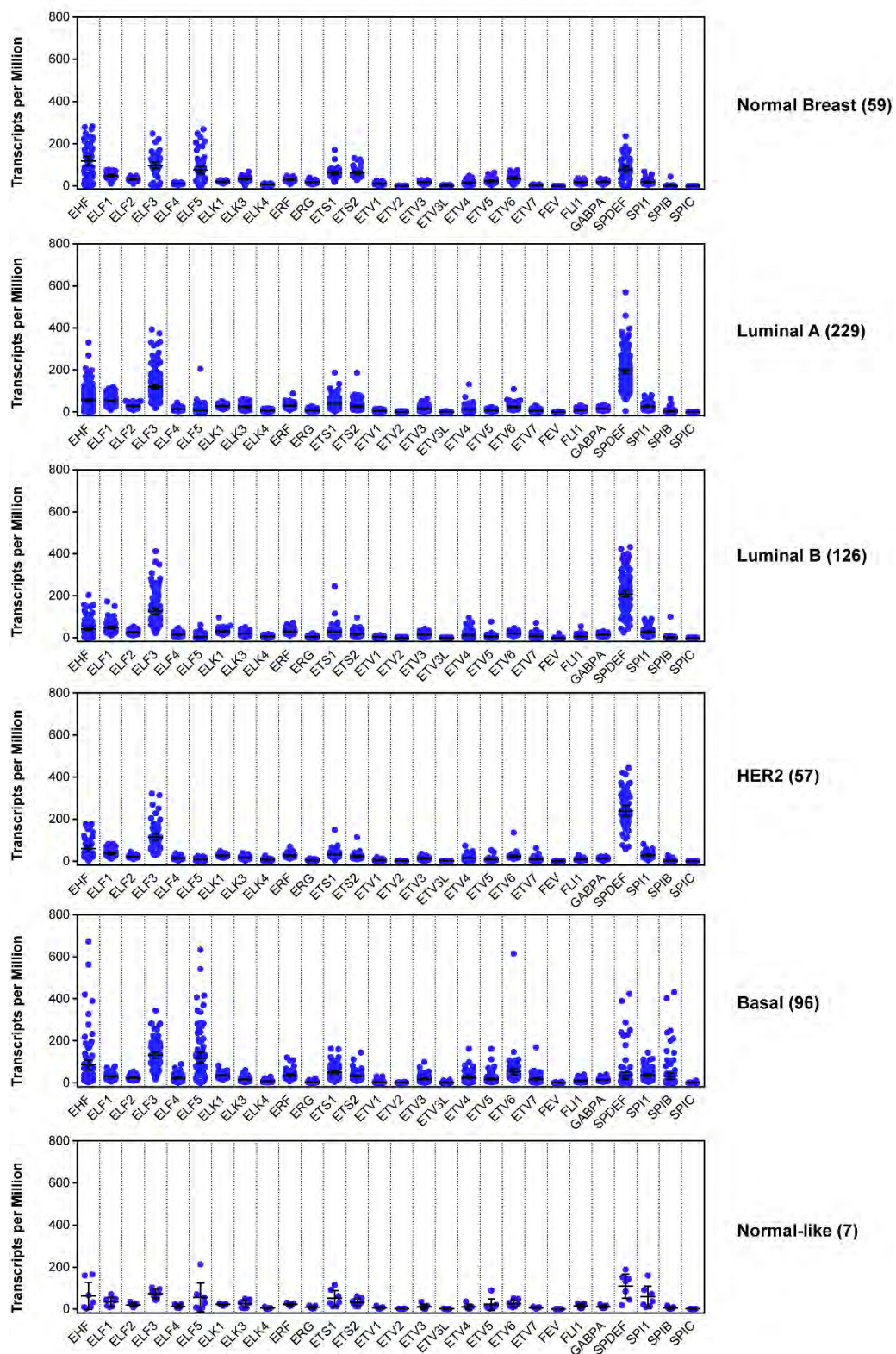
(A) *ELF5* gene expression (quantile normalised counts) for each breast cancer subtype plotted against American Joint Committee on Cancer (AJCC) TNM stage. Plotted values represent TCGA RNA-sequencing samples. (B) *ELF5* gene expression (quantile normalised counts) for each breast cancer subtype plotted against AJCC tumour stage. Plotted values represent TCGA RNA-sequencing samples.



**Figure 3.8: *ELF5* isoform expression in normal tissues and cancer (TPM values)**

(A-D) Equivalent to graphs shown in Figure 3.4B-D and Figure 3.6A using TPM values instead of quantile normalised counts. Plotted values represent individual samples and error bars show the mean with 95% confidence interval.





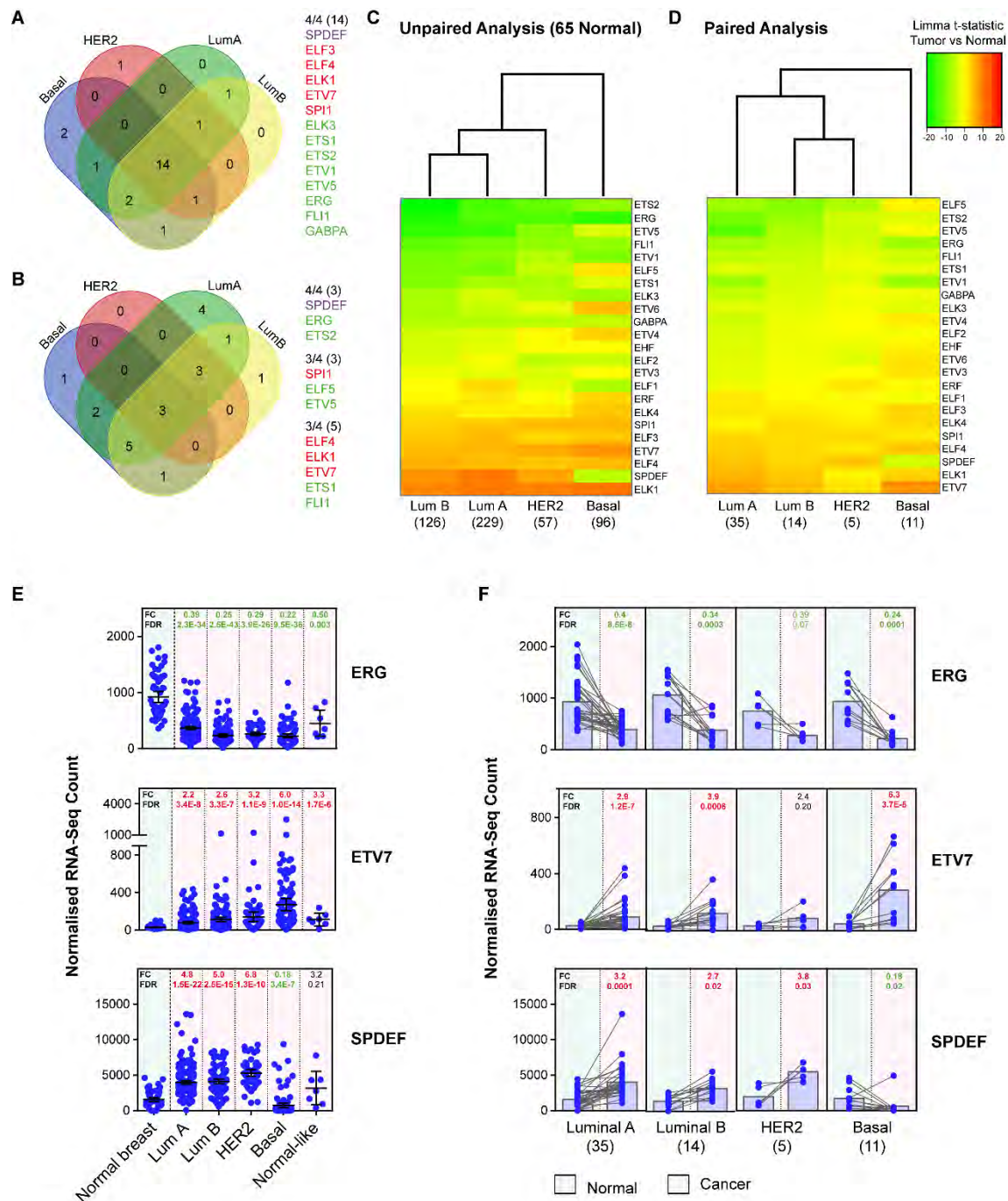
**Figure 3.9: ETS family gene expression in normal breast and breast cancer (TPM)**

Values are shown as 'Transcripts per million' (TPM), corrected for transcript length and allowing for limited comparison of expression within sample groups. Plotted values represent individual samples and error bars show the mean with 95% confidence interval. Numbers in parentheses after graph titles show the number of samples per group.

Of the 25 ETS factors included in the unpaired analysis, 24 were significantly altered in at least one subtype, with 14 common to all subtypes (Figure 3.10A). Within these, 13 were altered in the same direction (5 up and 8 down in the tumour compared to normal), while *SPDEF* was oppositely regulated in basal compared to other subtypes. In the paired analysis, 21 ETS factors were significantly altered in at least one subtype, with 3 ETS factors common to all subtypes (*SPDEF*, *ERG* and *ETS2*) and an additional 8 common to 3 of 4 subtypes (Figure 3.10B). *ELF5* was the most downregulated ETS family member by fold change in the luminal A, luminal B and HER2 subtypes in both unpaired and paired analyses.

Compared to other subtypes, the basal group showed a number of unique ETS factor expression changes. To further explore this, the limma t-statistic for all ETS family members (tumour compared to normal) were plotted on a clustered heatmap (Figure 3.10C, unpaired, and 3.10D, paired). The basal subtype showed a distinct expression profile and clustered separately from the other subtypes in both paired and unpaired analyses, highlighting the potential for the ETS transcription factor family to exert a unique transcriptional influence in this subtype. Similar results were obtained with unpaired and paired edgeR analyses (Additional Figure 3.2).

Several ETS family members with significant changes in expression were selected to visualise the results of the breast cancer differential expression analyses. The normalised counts for *ERG* (downregulated), *ETV7* (upregulated) and *SPDEF* (differentially regulated) are shown in Figure 3.10E. Direct comparison of matched normal and tumour samples is shown in Figure 3.10F. Interestingly, *SPDEF* showed the inverse expression pattern of *ELF5*. The normalised counts for the entire ETS factor family, with the results of the limma voom and edgeR differential expression analysis, are shown in Figure 3.11. TPM (within sample normalised) values for all ETS factors in breast cancer subtypes are shown in Figure 3.9.



**Figure 3.10: Expression of other ETS family members is also altered in breast cancer, with the basal subtype having a distinct ETS expression profile**

TCGA RNA-seq limma voom differential expression analysis data for ETS family members. (A) Venn diagram showing number of ETS family members significantly altered in breast cancer subtypes compared to normal (FDR<0.05). All subtypes were compared to a common set of 65 normal samples (unpaired analysis). Genes altered in all 4 subtypes are listed (red = upregulation, green = downregulation, purple = differentially regulated in basal subtype compared to other subtypes). (B) Venn diagram showing number of ETS family members significantly altered in breast cancer subtypes compared to normal (FDR<0.05), using paired normal and tumour samples from the same patient. Genes altered in at least 3

of 4 subtypes are listed, with colour-coding as above. (C) Clustered heat map of ETS factor limma voom t-statistic, comparing tumour samples to 65 normal samples. Legend is shown next to panel D. Rows are sorted by Luminal B values (smallest to largest) and columns are sorted according to clustering. Numbers in parentheses are samples per group.

(D) Clustered heat map of limma voom t-statistic, comparing paired normal and tumour samples, with sorting as above. Numbers in parentheses are sample pairs per group.

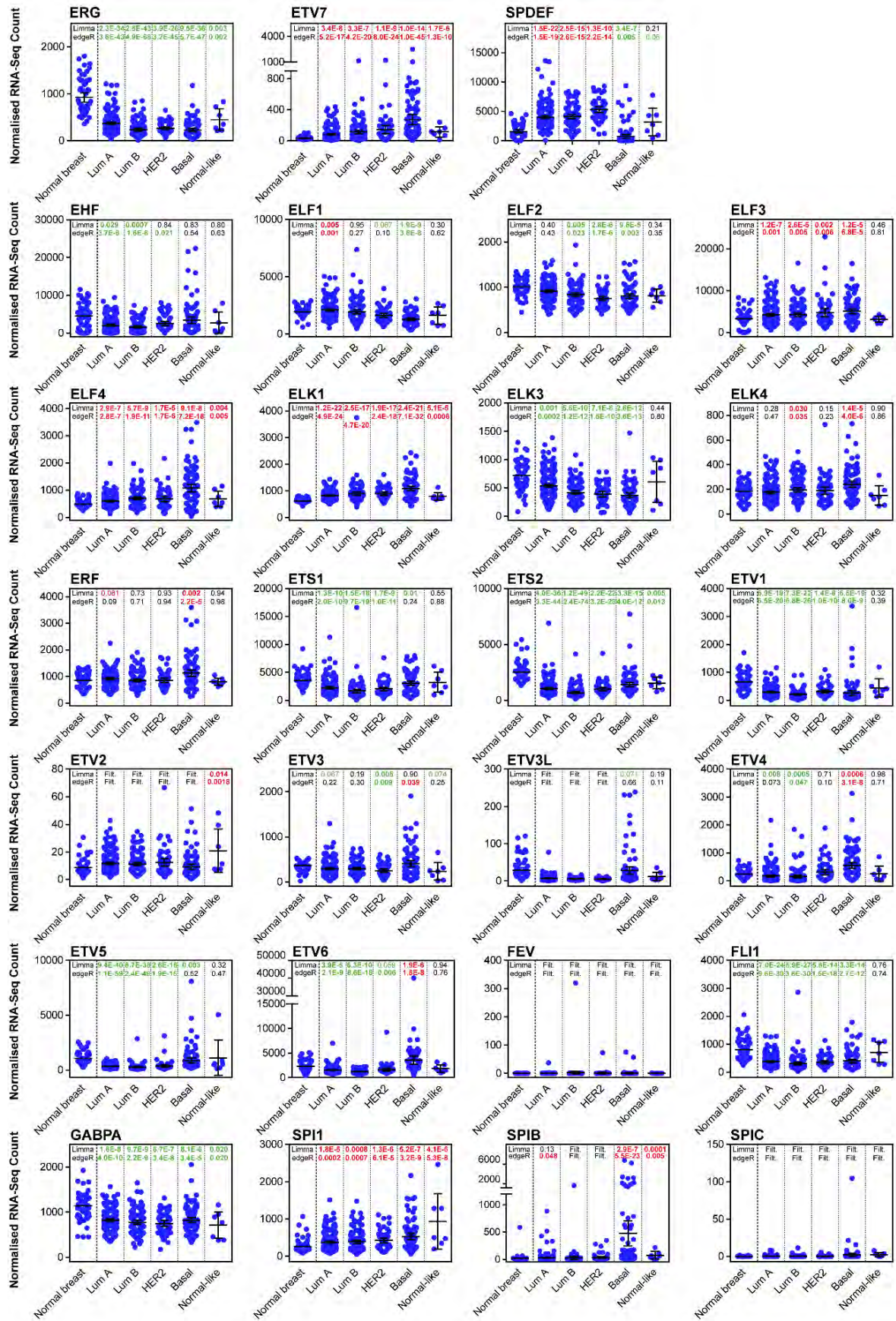
(E) Expression of *ERG*, *ETV7* and *SPDEF* for normal breast (green background) and breast cancer subtypes (pink background). Plotted values represent individual samples (normalised counts) and error bars the mean with 95% confidence interval. Fold change (FC) and False Discovery rate (FDR) from unpaired limma voom differential expression analysis are shown, with green indicating a significant downregulation and red a significant upregulation compared to normal (FDR<0.05). (F) *ERG*, *ETV7* and *SPDEF* levels for 65 patients with matched normal and cancer samples. Fold change (FC) and False Discovery rate (FDR) from paired limma voom differential expression analysis are shown, with colour-coding as above (FDR<0.05). Numbers in parentheses are sample pairs per group.

**Figure 3.11: ETS family expression gene expression in normal breast and breast cancer subtypes (normalised counts)**

**(next page)**

Values are shown as quantile normalised RNA-seq counts, allowing for comparison across subtypes but not correcting for transcript length. Plotted values represent individual samples and error bars show the mean with 95% confidence interval. False Discovery rates (FDR) from unpaired limma voom (top) and edgeR (bottom) differential expression analysis are shown, with bold green indicating a significant downregulation and bold red a significant upregulation compared to normal (FDR<0.05). Non-bold red or green values indicate FDR 0.05-0.10.

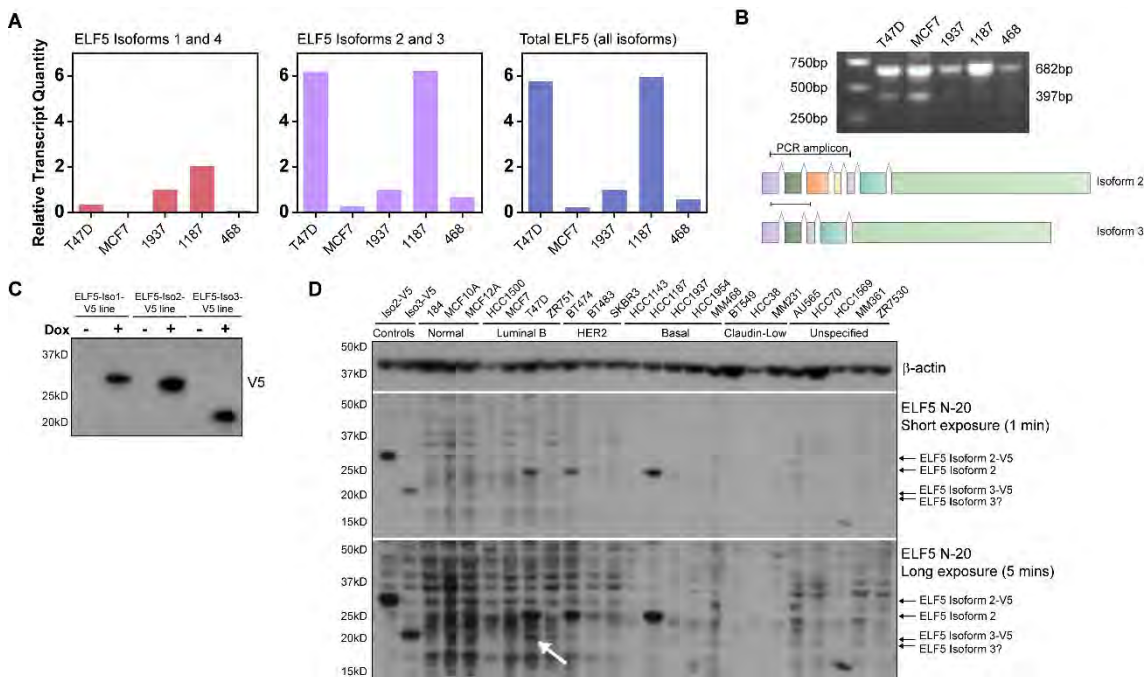




**Alterations in cell line *ELF5* isoform levels result in a similar phenotype, characterised by decreased cell number, decreased oestrogen-related proteins and nuclear localisation**

TCGA data showed an increased diversity of *ELF5* isoform expression in cancer compared to the normal breast. Therefore, the expression levels and effects of *ELF5* isoform expression were examined *in vitro* to determine if this was of functional relevance.

*ELF5* expression in a panel of breast cancer cell lines was analysed by qPCR (Figure 3.12A), end-point PCR (Figure 3.12B) and western blot (Figures 3.12 C and 3.12D). The PCR studies indicated that all 5 cell lines expressed *ELF5*, with expression pattern in the Isoform 2/3 assay closely resembling that seen for total *ELF5* (Figure 3.12A). Three cell lines (T47D, BT474 and HCC1187) expressed relatively high levels of *ELF5* protein, with the size of the main band consistent with Isoform 2. A possible band representing Isoform 3 was seen in the HCC1187 cell line, however interpretation was difficult due to high background (Figure 3.12D).



**Figure 3.12: *ELF5* mRNA and protein expression in breast cancer cell lines**

(A) qPCR for breast cancer cell lines showing relative levels of *ELF5* isoform pairs 1/4 and 2/3 and total *ELF5* (all isoforms). Single experiment with values normalised to HCC1937 samples. Abbreviations: 1937 = HCC1937, 1187 = HCC1187, 468 = MDA-MB-468. MDA-MB-231 cells have undetectable *ELF5* (not shown). (B) End-point PCR designed to amplify Isoforms 2 and 3 simultaneously in same panel of cell lines. DNA gel shows amplicons

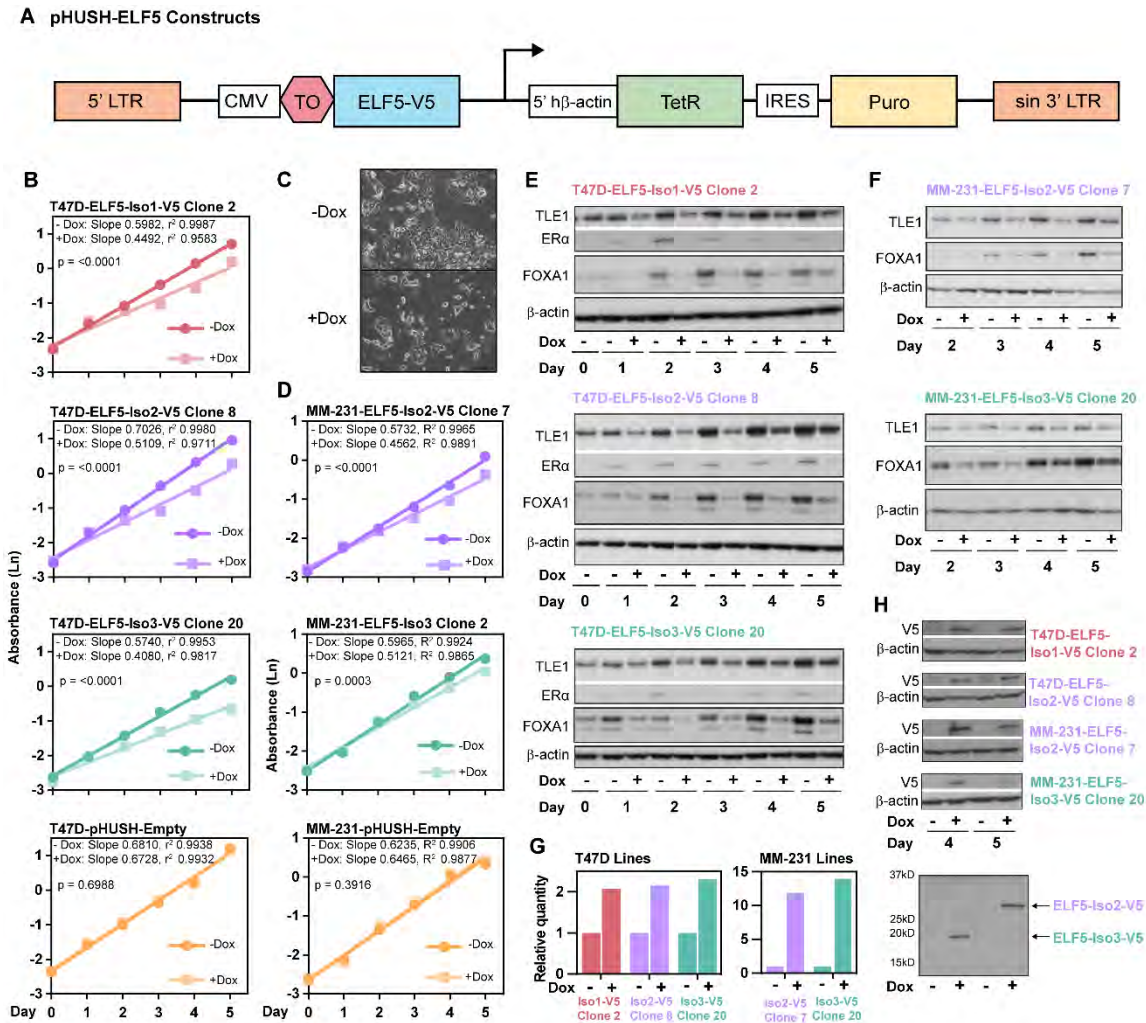
present after 25 PCR cycles. (C) V5 western blot of cell lines overexpressing ELF5 Isoform 1, 2 or 3 (tagged with V5) on addition of doxycycline (Dox), demonstrating relative isoform sizes. (D) Western blot for endogenous ELF5 in a panel of breast cancer cell lines, classified by molecular subtype. Controls in lanes 1 and 2 are cell lines overexpressing ELF5 Isoform 2 or 3 (tagged with V5) on addition of doxycycline (Dox). A possible ELF5 Isoform 3 band in HCC1187 cells is marked with a white arrow.

Clonal cell lines were constructed with a doxycycline-inducible expression vector containing a single ELF5 isoform, tagged with C-terminal V5 (Figure 3.13A). The luminal cell line T47D (ER+/PR+/HER2-) was chosen to examine the effect of isoforms in the context of relatively high endogenous ELF5 expression, testing the hypothesis that isoforms lacking the Pointed domain might exert a dominant-negative effect on full-length isoform function. A second Claudin-low line, MDA-MB-231 (ER-/PR-/HER2-), was chosen as it expresses no endogenous ELF5, allowing the effects of each isoform to be determined in the absence of potential competitive isoform interactions.

Over a 5-day timecourse, induced expression of Isoforms 1, 2 and 3 all resulted in a significantly decreased growth rate in T47D cells, with no change in the empty vector control (Figure 3.13B). Representative light microscope images for T47D lines (Figure 3.13C) demonstrate decreased cell number and increased detached cells; larger additional images shown in Figure 3.14A. A similar but less pronounced decrease in growth rate was also seen with induction of Isoform 2 and Isoform 3 in the MDA-MB-231 lines (Figure 3.13D and Figure 3.14B). It has been previously shown that the mechanisms underlying this phenotype for ELF5 Isoform 2 include G1 arrest, increased apoptosis and reduced adhesion proteins (Kalyuga *et al.*, 2012).

In the T47D lines, each isoform caused a decrease in oestrogen receptor-alpha (ER) protein and pioneer factors Forkhead Box A1 (FOXA1) and Transducin-like Enhancer of Split 1 (TLE1), required for ER-chromatin interactions (Holmes *et al.*, 2012; Hurtado *et al.*, 2011) (Figure 3.13E). The effects on FOXA1 and TLE1 were also seen in the MDA-MB-231 lines, in the absence of detectable ER (Figure 3.13F). Doxycycline-inducible *ELF5* mRNA expression was shown by qPCR (day 5, Figure 3.13G). V5 antibody western blot confirmed ELF5-V5 protein expression and also illustrated the size difference between Isoforms 2 and 3 (Figure 3.13H).

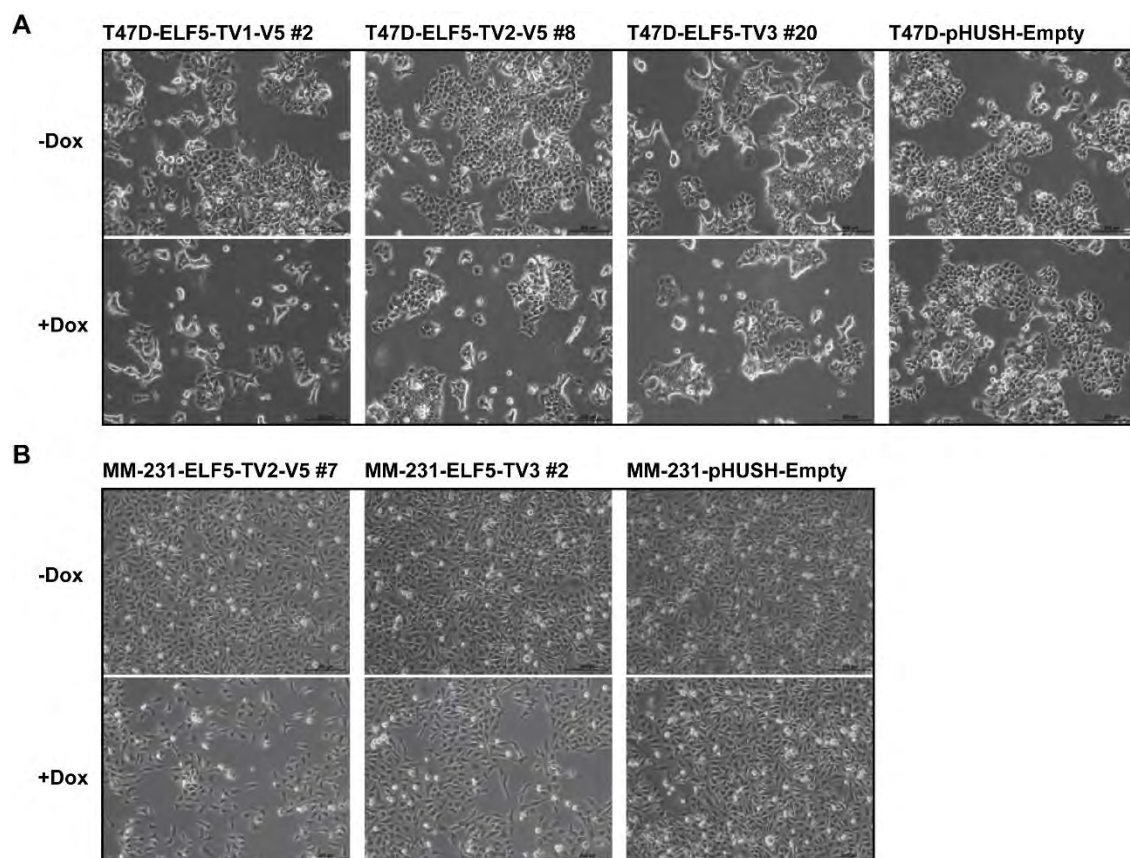




**Figure 3.13: Alterations in cell line ELF5 isoform levels result in a similar phenotype, characterised by decreased cell number and decreased expression of oestrogen-related proteins**

(A) Structure of the pHUSH-ProEx ELF5 retroviral expression vectors. In the absence of doxycycline, ELF5 isoform expression is inhibited by the tetracycline repressor (TetR). Doxycycline binds TetR, removing it from the Tet operon (TO) and allowing cytomegalovirus (CMV) promoter-driven expression of ELF5. TetR expression is linked to puromycin resistance (Puro) by an internal ribosome entry site (IRES). (B) and (D) Timecourse of T47D (B) and MDA-MB-231 (D) pHUSH clonal cell line growth +/- doxycycline over 5 days. Graphs show the natural logarithm (Ln) of spectrophotometric assay absorbance value (y-axis) plotted against day (x-axis). p-values compare -dox and +dox slopes for each cell line. One experiment shown. (C) Representative light microscope images of T47D cells +/- doxycycline, taken at day 4. (E) Western blots for oestrogen-related proteins from T47D timecourses, days 0-5. (F) Western blots for oestrogen-related proteins from MDA-MB-231 timecourses, days 2-5. (G) qPCR for ELF5 (day 5 timecourse samples) +/- doxycycline. (H) Western blots for V5 at days 4 and 5 +/- doxycycline, 65ug per lane (T47D-ELF5-Isoform2-V5 line) or 25ug per lane (all others). Bottom panel shows representative samples from MDA-MB-231 cell lines, demonstrating the size difference between Isoforms 2 and 3.

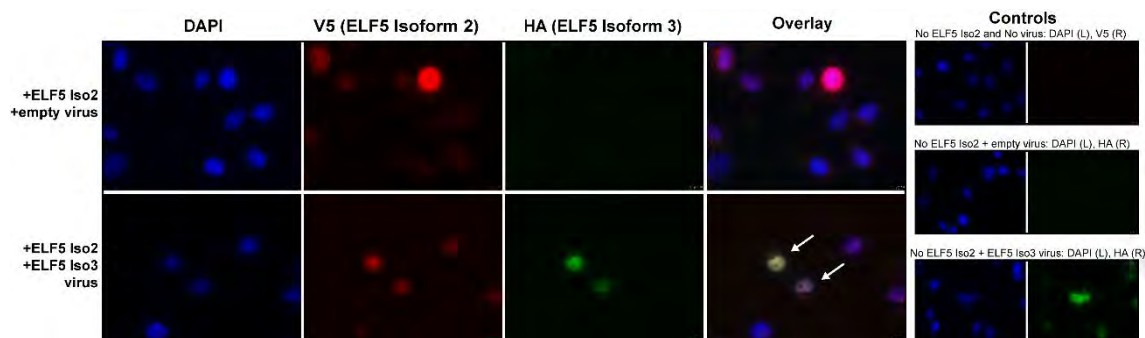




**Figure 3.14: Phenotype of pHUSH-ELF5-V5 breast cancer clonal cell lines**

Light microscope images taken at day 4 for T47D clonal cell lines (A) and MDA-MB-231 cell lines (B), treated with vehicle (top row) or doxycycline (bottom row).

Immunofluorescence was performed to determine the subcellular location of ELF5 isoforms when expressed in isolation and when co-expressed. MDA-MB-231 cells with doxycycline-inducible ELF5-Isoform 2-V5 expression were used, with transient retroviral infection of an ELF5-Isoform 3-HA vector; this allowed manipulation of Isoform 2 and Isoform 3 levels within the same cell. Figure 3.15 (top row) shows MDA-MB-231-ELF5-Isoform 2-V5 cells treated with doxycycline to induce expression, as well as transient infection of a control pQCXIH vector; there was strong nuclear V5 staining and no HA staining. In row 2, cells were treated with doxycycline to induce ELF5-Isoform 2-V5 and also infected with Isoform 3-HA. Both Isoform 2 (V5) and Isoform 3 (HA) localised to the nucleus and there was no cytoplasmic redistribution seen in the cells that expressed both Isoform 2 and Isoform 3 (indicated by arrows), an effect which has been reported previously for ETS1 isoforms (Laitem *et al.*, 2009).



**Figure 3.15: ELF5 Isoform 3 expression does not alter Isoform 2 subcellular localisation**

Immunofluorescent images of MDA-MB-231-ELF5-Isoform 2-V5 Clone 7 cells. Cells were treated with doxycycline to induce ELF5-Isoform2-V5 expression, and infected with pQCXIH-empty-vector (control) or pQCXIH-ELF5-Isoform3-HA retrovirus. Negative controls (-/+ retrovirus but no doxycycline treatment) are shown on the right. Blue = nuclei (DAPI), red = V5 (ELF5 Isoform 2), green = HA (ELF5 Isoform 3). Arrows mark cells with double Isoform 2/3 expression.

### ELF5 isoforms have a similar transcriptional effect in T47D and MDA-MB-231 cell lines

A panel of 116 genes was examined by qPCR to compare the transcriptional effects of *ELF5* isoforms. Previously published microarrays and ELF5/V5 chromatin immunoprecipitation sequencing (ChIP-seq) (Kalyuga *et al.*, 2012) were used to identify genes and pathways regulated by ELF5 Isoform 2 in luminal cell lines. An outline of the experimental workflow is shown in Figure 3.16.

The pHUSH clonal cell lines were selected based on similar qPCR levels of *ELF5* isoform induction. Figure 3.17A shows the *ELF5* level +dox relative to the -dox control for each individual cell line. To compare baseline (-dox) variability, values were also normalised to the lowest ELF5 value (Figure 3.17B). Baseline variability was minimal in the T47D lines, however ranged from 1.0-2.3x in the MDA-MB-231 Isoform 3 lines and 4.7x (clone 6) to 28.0x (clone 1) in the Isoform 2 lines. This variation is most likely due to slight “leakiness” of the pHUSH vector, leading to low-level ELF5 expression (undetectable by V5 western blot) in the absence of doxycycline.

T47D and MDA-MB-231 clonal cell lines were treated with doxycycline or vehicle for 48 hours to induce ELF5 isoform expression. Initially, 2 clones per parental cell line were used. A selection of 27 genes was then repeated in 1 or 2 further clones, giving a total of 3-4 clonal lines (biological replicates) per parental line (Table 3.3). The heat maps in

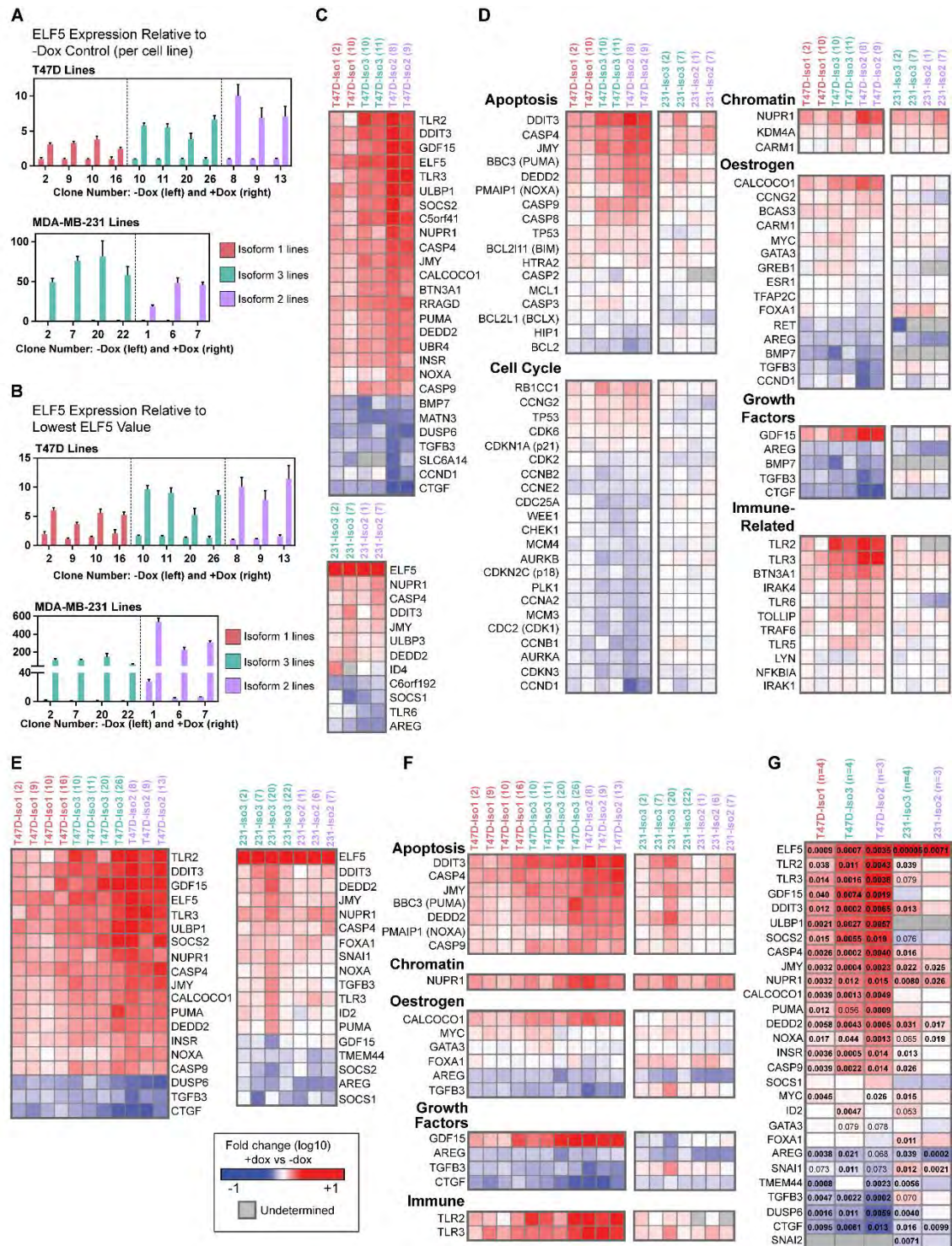


Overall, the pattern of change was fairly similar regardless of which ELF5 isoform was expressed. The genes with the strongest absolute fold change ( $>3$  in any T47D line or  $>2$  in any MDA-MB-231 line) showed a particularly consistent pattern of change (Figure 3.17C). Expression changes were greater in the T47D than in the MDA-MB-231 cell lines.

Genes were also analysed in functional categories (Figure 3.17D). Apoptosis-related genes showed consistent changes corresponding to an increase in apoptosis, for example upregulation of apoptosis-promoting genes such as *DDIT3*, *PUMA*, *NOXA*, *TP53* and various caspases, as well as downregulation of apoptosis-inhibiting genes such as *BCLX* and *BCL2*. The changes in cell cycle genes were weaker, although still generally consistent, with up-regulation of cell cycle inhibitors such as *RB1CC1* and *TP53* and down-regulation of cell cycle promoting genes such as cyclins D1, B1, A2, E2 and associated kinases *CDK1/2*. However, the pattern of change was not entirely congruent with inhibition of the cell cycle, with upregulation of the cyclin-D-associated *CDK6* and downregulation of the cell cycle inhibitor *CDKN2C* (p18). Changes in mRNA expression for key genes associated with oestrogen action such as *ESR1*, *FOXA1*, *GATA3* and *GREB1* were relatively small and variable (Figure 3.17D), in contrast to results at the protein level, which showed robust downregulation of ER and FOXA1 with all ELF5 isoforms.

The results were substantiated using 1-2 further clones per parental cell line and 27 genes from the original panel (Figures 3.17E and 3.17F). The average fold change for each parental cell line group (consisting of 3-4 clonal cell lines) was calculated and this is shown in the heat map in Figure 3.17G with corresponding significant p-values (FDR  $<0.10$ ). Although the pattern of change was generally consistent, there were some interesting differences. Firstly, *FOXA1* expression in the T47D lines exhibited a mostly downward trend, although there were no statistically significant changes. Conversely, in the MDA-MB-231 lines *FOXA1* mRNA increased (significant only in the Isoform 2 group); again, this is in contrast to the protein results shown for the MDA-MB-231 lines in Figure 3.13F. Secondly, there was only one case in the T47D lines (and none in the MDA-MB-231 lines) in which a gene was altered in statistically significant opposite directions by different ELF5 isoforms. This gene, *GATA3*, was upregulated by Isoform 3 and downregulated by Isoform 2, although the changes were relatively small. In fact, 20 of 27 genes in the T47D lines showed a statistically significant change in the same direction with each of the 3 isoforms, pointing towards the overall consistency of the transcriptional effect of ELF5 isoforms.





**Figure 3.17: ELF5 isoforms have a similar transcriptional effect in T47D and MDA-MB-231 cell lines**

(A) *ELF5* expression measured by qPCR at 48 hours for T47D clonal cell lines (top) and MDA-MB-231 clonal cell lines (bottom). Assay detects all *ELF5* isoforms. Values are the mean Calibrated Normalised Relative Quantity (CNRQ) with standard error. Results relative to the -dox control (set at 1) for each cell line. (B) *ELF5* expression measured by qPCR at 48 hours for T47D clonal cell lines (top) and MDA-MB-231 clonal cell lines (bottom) used in

the qPCR panel. Assay detects all *ELF5* isoforms. Values are the mean CNRQ with standard error. Results relative to the sample with the lowest *ELF5* value (set at 1), which is T47D-*ELF5*-Isoform 2-V5 Clone 8 (T47D lines) and MDA-MB-231-*ELF5*-Isoform 3-V5 Clone 22 (MDA-MB-231 lines). (C) Heat map showing genes (from 116-gene qPCR panel) with absolute fold change >3 (any T47D line) or >2 (any MDA-MB-231 line). 2 clonal cell lines tested per group. All heat maps use the legend shown in panel E and represent the log<sub>10</sub> fold change (capped at -1 and +1) of the +dox quantity compared to the -dox quantity as measured by qPCR. Grey indicates gene was not detectable by qPCR in -dox and/or +dox samples. (D) Functional categorisation of selected genes from 116-gene qPCR panel. Some genes are represented more than once due to multiple functions. (E) Heat map showing genes (from 27-gene qPCR panel) with absolute fold change >3 (any T47D line) or >2 (any MDA-MB-231 line). Results shown for 3-4 clonal lines per group. (F) Functional categorisation of selected genes from 27-gene qPCR panel. (G) Heat map representing the mean log<sub>10</sub> fold change per group for all genes in the 27-gene panel (+*ELF5*). Significant p-values are shown where false discovery rate (FDR) <0.10. Some p-values (non-bold) are >0.05, although FDR for these values is <0.10. Non-significant p-values (FDR >0.10) are not shown.

**Table 3.3: Clonal cell lines used in qPCR panel**

	Isoform 1	Isoform 2	Isoform 3	Isoform 4	Isoform 5
Parental line	T47D-pHUSH-ELF5-Isoform 1-V5 (pool)	T47D-pHUSH-ELF5- Isoform 2-V5 (pool)	T47D-pHUSH-ELF5- Isoform 3-V5 (pool)		
T47D clones qPCR round 1	T47D-pHUSH-ELF5-Isoform 1-V5 Clone 2*	T47D-pHUSH-ELF5- Isoform 2-V5 Clone 8*	T47D-pHUSH-ELF5- Isoform 3-V5 Clone 10	Not tested	Not tested
	T47D-pHUSH-ELF5- Isoform 1-V5 Clone 10	T47D-pHUSH-ELF5- Isoform 2-V5 Clone 9	T47D-pHUSH-ELF5- Isoform 3-V5 Clone 11		
T47D clones qPCR round 2	T47D-pHUSH-ELF5- Isoform 1-V5 Clone 9	T47D-pHUSH-ELF5- Isoform 2-V5 Clone 13	T47D-pHUSH-ELF5- Isoform 3-V5 Clone 20*		
	T47D-pHUSH-ELF5- Isoform 1-V5 Clone 16		T47D-pHUSH-ELF5- Isoform 3-V5 Clone 26		
Parental line		MDA-MB-231-pHUSH-ELF5-TV2-V5 (pool)	MDA-MB-231-pHUSH-ELF5-Isoform 3-V5 (pool)		
MDA-MB-231 clones qPCR round 1	Not tested	MDA-MB-231-pHUSH-ELF5-Isoform 2-V5 Clone 1	MDA-MB-231-pHUSH-ELF5-Isoform 3-V5 Clone 2*	Not tested	Not tested
		MDA-MB-231-pHUSH-ELF5-Isoform 2-V5 Clone 7*	MDA-MB-231-pHUSH-ELF5-Isoform 3-V5 Clone 7		
MDA-MB-231 clones qPCR round 2		MDA-MB-231-pHUSH-ELF5-Isoform 2-V5 Clone 6	MDA-MB-231-pHUSH-ELF5-Isoform3-V5 Clone 20		
			MDA-MB-231-pHUSH-ELF5-Isoform 3-V5 Clone 22		

All clonal lines were derived from a parental line as listed. Clones were either used in round 1 (116 genes) or round 2 (27 genes). Asterisk (\*) indicates that line was also used in the timecourse experiment shown in Figure 3.13.

## Discussion

This study is the first detailed analysis of ELF5 isoform expression and function, extending previous ELF5 Northern blotting, immunohistochemistry and microarray studies (Kalyuga *et al.*, 2012; Lapinskas *et al.*, 2004; Oettgen *et al.*, 1999; Zhou *et al.*, 1998) to the isoform level using 6,757 sequenced normal and cancer samples. The kidney appears to be unique in being the only tissue examined to express Isoform 1 as its dominant isoform, expanding on the initial Northern blot descriptions of ELF5 isoforms (Oettgen *et al.*, 1999). In breast cancer, ELF5 alterations were subtype-specific, with the basal subtype demonstrating unique ELF5 isoform expression changes. Despite differences in protein domains, the *in vitro* phenotypic and transcriptional effects of increased ELF5 isoform expression were similar. This suggests that ELF5 action is regulated in various tissues by tissue-specific alternative promoter use rather than by differences in the transcriptional activity of the isoforms.

In cancer, *ELF5* expression is frequently altered. The kidney, one of the highest *ELF5*-expressing tissues, showed a dramatic decrease in *ELF5* level in cancer. ELF5 has been characterised as a tumour suppressor in the kidney and bladder (Lapinskas *et al.*, 2011; Wu *et al.*, 2015) and this may restrict kidney carcinomas to non-*ELF5*-expressing cells of origin. In other tissues, cancer was associated with an aberrant increase in *ELF5* expression, as seen in the cervix, colon, rectum and uterus. This may indicate an oncogenic role for ELF5 in these tissues or broader genomic deregulation, for example DNA hypomethylation, a hallmark of the cancer genome (Gama-Sosa *et al.*, 1983). The mechanisms regulating ELF5 in different tissues and in cancer have not been widely studied, however in the early embryo and the developing mammary gland, ELF5 regulation of lineage specification is associated with promoter methylation status (Lee *et al.*, 2011a; Ng *et al.*, 2008). Increased ELF5 promoter methylation has also been demonstrated in bladder carcinoma (Wu *et al.*, 2015). These studies establish DNA methylation as an important epigenetic mechanism regulating ELF5 expression, with possible aberrant methylation in cancer.

The normal human breast expresses relatively high levels of *ELF5*, with subtype-specific alterations in cancer. High ELF5 has been shown to maintain the ER-negative basal phenotype, paralleling the normal developmental role of specification of the ER-negative alveolar lineage (Kalyuga *et al.*, 2012). In all breast cancer subtypes, there was a broader distribution of *ELF5* isoform expression. Increased variability of isoform distribution (“transcriptome instability”) is a known phenomenon and is proposed as a



molecular hallmark of cancer (Sveen *et al.*, 2014; Venables *et al.*, 2009). A recent study identified 244 cancer-associated isoform “switches”, involving consistent changes in the most abundant isoform (Sebestyen *et al.*, 2015). An ELF5 isoform “switch” has not been identified in breast cancer, in keeping with the current study, which showed an inconsistent pattern of isoform expression variation. Although not consistently identified, this does not mean that ELF5 isoform switches may not be playing an important role in the subset of patients in which they occur.

Other ETS transcription factors have also shown to be important in breast cancer and extension of RNA-sequencing analysis to the entire ETS family revealed a number of cancer-associated expression changes. The ETS family as a whole has been previously studied in breast cancer at the qPCR level in mouse models (Galang *et al.*, 2004) and human cell lines (He *et al.*, 2007), although this is the first study to examine the expression of the entire human ETS family in both the normal breast and subtyped breast cancer samples using RNA-sequencing data. The normal human breast expressed a diverse range of ETS factors. Compared to the normal breast, the basal-like subtype showed a distinct pattern of ETS factor expression changes, with several ETS factors changing in the opposite direction in basal compared to other subtypes. *ELF5* and *SPDEF* were the most striking examples of this. *SPDEF* is also a luminal epithelial lineage-specific transcription factor in the breast and has been shown to promote the survival of ER-positive breast cancer cells (Buchwalter *et al.*, 2013). The inverse relationship seen between these 2 transcription factors in breast cancer is intriguing and may well have a parallel during normal mammary development.

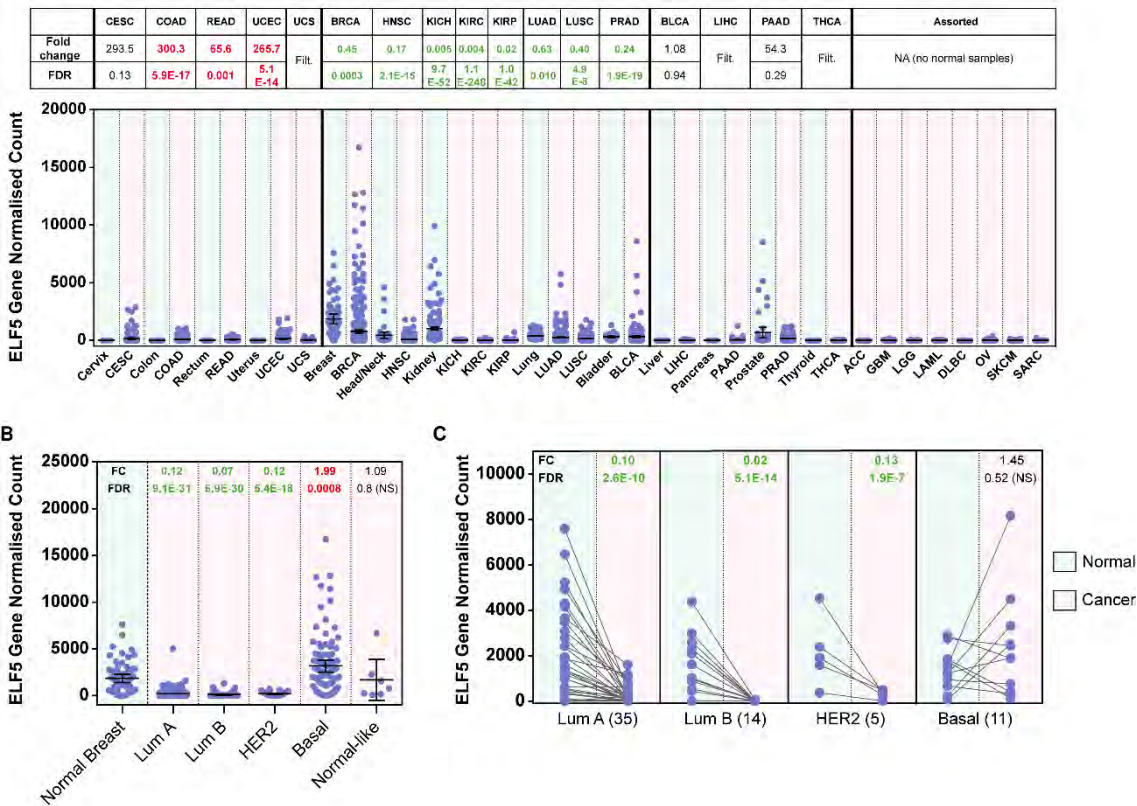
Finally, the phenotypic and transcriptional effects of Isoforms 1, 2 and 3 were found to be similar in inducible cell line models. This was unexpected, as the Pointed domain in murine ELF5 has been previously shown to have strong transactivation activity (Choi and Sinha, 2006). In many proteins, SAM/PNT domains act as protein-protein interaction modules, an important mechanism of biological specificity for ETS factors, which often bind only weakly to DNA in the absence of binding partners or post-translational modifications (Choi and Sinha, 2006; Li *et al.*, 2000). The importance of the Pointed domain is also shown by other ETS family members in which removal of the Pointed domain significantly alters protein function. The endogenous ETS1 isoform p27, for example, lacks the Pointed and transactivation domains and negatively regulates full-length ETS1 by competing for DNA binding sites and promoting its translocation from the nucleus to the cytoplasm (Laitem *et al.*, 2009). Although this splicing event is similar to those that occur to produce ELF5 isoforms 3 and 4, it

appears that ELF5 Isoform 3 can alter gene transcription in a very similar way to the full-length isoforms. The T47D cell line (with relatively high endogenous ELF5 expression) was selected to study the potential competitive effects of ELF5 Isoform 3, while the MDA-MB-231 cell line (undetectable endogenous ELF5 expression) was chosen to investigate the transcriptional effects of ELF5 Isoform 3 in the absence of competition. Future studies with other cell lines (for example, a luminal breast cancer cell line with low endogenous ELF5 expression such as MCF7) may help to clarify the role of ELF5 Isoform 3 in breast cancer cells.

In addition, there was no subcellular relocation of full-length Isoform 2 seen when Isoform 3 was co-expressed. Interestingly, however, while exogenous ELF5 localised to the nucleus in this study, cytoplasmic ELF5 staining is seen in some human breast cancer samples and is a predictor of outcome (Gallego-Ortega *et al.*, 2015). This indicates that endogenous ELF5 can localise to the cytoplasm and that this has functional significance in breast cancer. A potential nuclear export sequence (NES) exists in the ETS domain of ELF5 (amino acids 165-174) similar to one identified in ELF3 (Prescott *et al.*, 2004; Prescott *et al.*, 2011). It is possible that cytoplasmic relocation of ELF5 is mediated by the relative amounts of isoforms but that this effect is not recapitulated by exogenous expression, particularly in the context of MDA-MB-231 cells which do not normally express ELF5 and therefore may be lacking essential protein binding partners. Given the importance of context in the function of ETS factors, it is possible that the differential effects of ELF5 isoforms may also require a stimulus (for example, growth factors) or challenge (for example, oestrogen deprivation) in order to become apparent, an avenue that was not explored in this study.

## Additional Figures

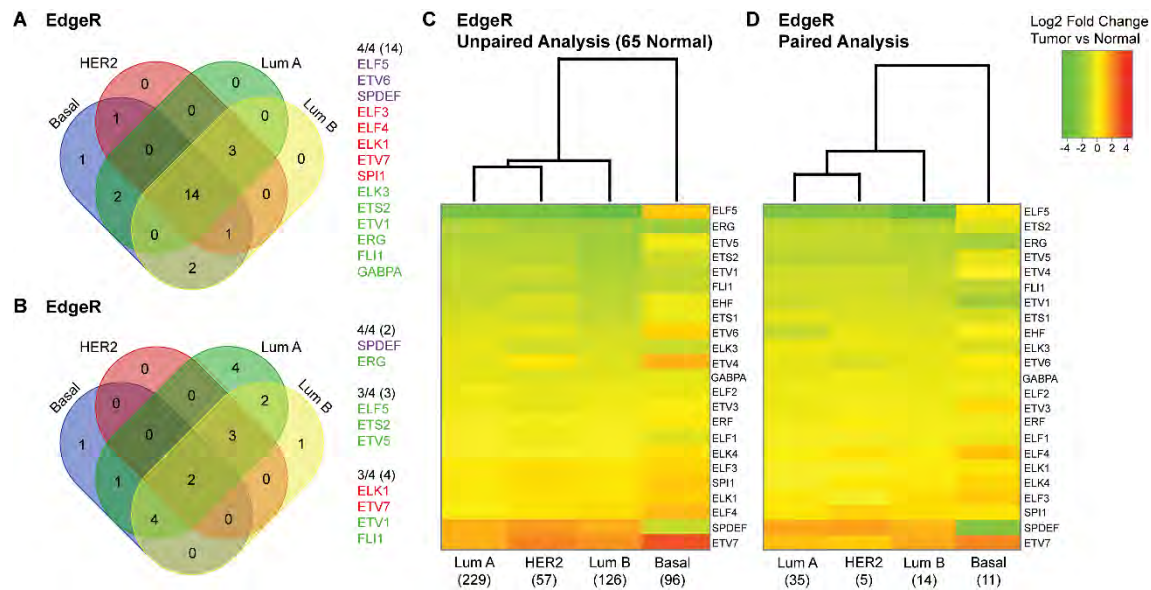
### A EdgeR differential expression analysis



**Additional Figure 3.1: *ELF5* expression is significantly altered in cancer - results from edgeR differential expression analysis**

All fold change (FC) and False Discovery rate (FDR) are from edgeR (instead of limma voom) differential expression analysis, with green values in bold indicating a significant downregulation and red values in bold a significant upregulation compared to normal (FDR<0.05). Filt. indicates gene filtered from edgeR analysis due to low expression.

(A) *ELF5* gene expression (quantile normalised counts) for selected normal tissues and cancers with edgeR FC and FDR values. (B) *ELF5* gene expression (quantile normalised counts) for normal breast and breast cancer subtypes, as a comparison for Figure 4A. (C) *ELF5* gene expression (quantile normalised counts) for patients with matched normal and cancer samples with edgeR FC and FDR values. Numbers in parentheses indicate sample pairs per group.



**Additional Figure 3.2: Expression of other ETS family members is also altered in breast cancer, with the basal subtype having a distinct ETS expression profile - results from edgeR differential expression analysis**

TCGA RNA-seq edgeR differential expression analysis data for ETS family members, using results from edgeR (instead of limma voom) differential expression analysis. (A) Venn diagram showing number of ETS family members significantly altered in breast cancer subtypes compared to normal (FDR<0.05). All subtypes were compared to a common set of 65 normal samples (unpaired analysis). Genes altered in all 4 subtypes are listed (red = upregulation, green = downregulation, purple = differentially regulated in basal subtype compared to other subtypes). (B) Venn diagram showing number of ETS family members significantly altered in breast cancer subtypes compared to normal (FDR<0.05), using paired normal and tumour samples from the same patient. Genes altered in at least 3 of 4 subtypes are listed, with colour-coding as above. (C) Clustered heat map of ETS factor edgeR log2 fold change, comparing tumour samples to 65 normal samples. Legend is shown next to panel D. Rows are sorted by Luminal B values (smallest to largest) and columns are sorted according to clustering. Numbers in parentheses are samples per group. (D) Clustered heat map of ETS factor edgeR log2 fold change, comparing paired normal and tumour samples, with sorting as above. Numbers in parentheses are sample pairs per group.



## Appendix: ELF5 isoform expression is tissue-specific and significantly altered in cancer (Piggin *et al*, 2016)

Piggin *et al. Breast Cancer Research* (2016) 18:4  
DOI 10.1186/s13058-015-0666-0

Breast Cancer Research

### RESEARCH ARTICLE

### Open Access



# ELF5 isoform expression is tissue-specific and significantly altered in cancer

Catherine L. Piggin<sup>1\*</sup> , Daniel L. Roden<sup>1</sup>, David Gallego-Ortega<sup>1</sup>, Heather J. Lee<sup>1,2</sup>, Samantha R. Oakes<sup>1</sup> and Christopher J. Ormandy<sup>1</sup>

### Abstract

**Background:** E74-like factor 5 (ELF5) is an epithelial-specific member of the E26 transforming sequence (ETS) transcription factor family and a critical regulator of cell fate in the placenta, pulmonary bronchi, and milk-producing alveoli of the mammary gland. ELF5 also plays key roles in malignancy, particularly in basal-like and endocrine-resistant forms of breast cancer. Almost all genes undergo alternative transcription or splicing, which increases the diversity of protein structure and function. Although ELF5 has multiple isoforms, this has not been considered in previous studies of ELF5 function.

**Methods:** RNA-sequencing data for 6757 samples from The Cancer Genome Atlas were analyzed to characterize ELF5 isoform expression in multiple normal tissues and cancers. Extensive *in vitro* analysis of ELF5 isoforms, including a 116-gene quantitative polymerase chain reaction panel, was performed in breast cancer cell lines.

**Results:** ELF5 isoform expression was found to be tissue-specific due to alternative promoter use but altered in multiple cancer types. The normal breast expressed one main isoform, while in breast cancer there were subtype-specific alterations in expression. Expression of other ETS factors was also significantly altered in breast cancer, with the basal-like subtype demonstrating a distinct ETS expression profile. *In vitro* inducible expression of the full-length isoforms 1 and 2, as well as isoform 3 (lacking the Pointed domain) had similar phenotypic and transcriptional effects.

**Conclusions:** Alternative promoter use, conferring differential regulatory responses, is the main mechanism governing ELF5 action rather than differential transcriptional activity of the isoforms. This understanding of expression and function at the isoform level is a vital first step in realizing the potential of transcription factors such as ELF5 as prognostic markers or therapeutic targets in cancer.

**Keywords:** ELF5, ETS transcription factors, Isoforms, Transcript variants, Splicing, Cancer

### Background

Transcription factors are the integrators of multiple signaling pathways, converting internal and external stimuli into changes in gene expression. Through this role, the evolutionarily conserved E26 transforming sequence (ETS) transcription factor family controls fundamental cellular processes such as proliferation, differentiation, and apoptosis [1]. The 28 members of the human ETS family are characterized by an ETS DNA-binding domain that recognizes a core GGAA/T motif. Additional specificity of ETS domain binding is conferred by the amino

acids surrounding the key residues, as well as by post-translational modifications and interactions with other proteins [2, 3]. Given the vital cellular processes regulated by ETS transcription factors, it is not surprising that they have also been identified as significant contributors to tumorigenesis [4].

E74-like factor 5 (ELF5) is an epithelial-specific member of the ETS transcription factor family [5, 6]. In addition to the ETS domain, the full-length ELF5 protein contains an N-terminal Pointed (PNT) domain (83 amino acids) that is similar to the evolutionarily conserved sterile alpha motif (SAM) domain. In humans, the SMART database [7] identifies 96 SAM/PNT domain-containing proteins, 11 of which are ETS family members. SAM domains have diverse functions, including protein–protein interactions,

\* Correspondence: c.piggin@garvan.org.au

<sup>1</sup>Cancer Division, Garvan Institute of Medical Research/The Kinghorn Cancer Centre, Sydney, NSW 2010, Australia

Full list of author information is available at the end of the article



© 2016 Piggin *et al.* **Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated.

polymerization, kinase docking, RNA binding, and lipid molecule interactions [8–11]. The ELF5 PNT domain has been shown to have strong transactivation activity [12]; however, the mechanisms underlying this activity (for example, protein–protein interactions or posttranslational modifications) are unknown.

A critical function of ELF5 is the regulation of cell fate, beginning with specification of the trophectoderm in the blastocyst [13]. Correct spatial and temporal ELF5 expression is also important for normal development of the embryonic lung [14]. In the mammary gland, prolactin- and progesterone-driven ELF5 expression during pregnancy directs the development of the luminal progenitor cells into estrogen receptor- $\alpha$  (ER)- and progesterone receptor (PR)-negative milk-producing cells [15]. In normal human tissues, ELF5 is reported to be expressed in the kidney, prostate, lung, mammary gland, salivary gland, placenta, and stomach [5, 6, 16].

More recently, there has been increasing interest in the role of ELF5 in cancer. ETS factors are frequently deregulated in cancer through diverse mechanisms, including gene fusions, alterations in localization and/or activity, amplifications, increased expression, and (less commonly described) decreased expression [4]. ELF5 was originally described as a tumor suppressor [5]; however, the role of this protein in cancer is complex and context-dependent. In prostate cancer, for example, ELF5 has been shown to inhibit transforming growth factor (TGF)- $\beta$ -driven epithelial–mesenchymal transition by blocking phosphorylation of the TGF- $\beta$  effector protein SMAD3 [17]. Conversely, *ELF5* mRNA has been shown to be upregulated in a cell line model of prostate cancer progression involving acquisition of androgen independence [18]. Bladder and kidney carcinoma have been associated with loss of ELF5 expression at the protein and RNA levels [19, 20], whereas in endometrial carcinoma *ELF5* upregulation is associated with higher disease stage [21]. *ELF5* gene rearrangements have been described in several lung cancer cell lines [5], and the authors of a recent case study described a *ZFPM2-ELF5* fusion gene in multicystic mesothelioma [22]; however, gene fusions do not appear to be a major mechanism for deregulation of ELF5, in contrast to other ETS factors, such as *TMPRSS2-ERG/ETV1* fusions in prostate cancer [23].

The breast is the most well-studied context for the role of ELF5 in cancer, with microarrays showing increased expression in basal-like subtypes and decreased expression in luminal A/B and Erb-b2 receptor tyrosine kinase 2 (HER2)-overexpressing subtypes [24, 25], suggesting subtype-specific effects. Transient ELF5 expression in cell line models reduced proliferation, invasion, ER-driven transcription and epithelial–mesenchymal transition [25, 26]. However, sustained increased ELF5 expression in some contexts is associated with disease progression, such as in

endocrine-resistant breast cancers, reliant on elevated ELF5 for growth in cell line models, and the basal-like subtype of breast cancer [25]. This illustrates the complexity and contextual dependence of transcriptional regulation.

It is becoming increasingly recognized that almost all multiexon genes undergo alternative transcription (such as alternative transcription start or termination sites) and/or alternative exon splicing, increasing diversity of protein structure and function [27]. Alternative transcription events are also commonly deregulated in cancer, contributing to tumor initiation and progression but also providing potential cancer-specific therapeutic targets. Importantly, different isoforms produced by the same gene may have very different functions. One striking example is vascular endothelial growth factor, which produces both proangiogenic and antiangiogenic isoforms [28]. Early studies described tissue-specific differences in *ELF5* transcript isoform expression [6], but recent studies have not distinguished between isoforms or have used a single isoform for overexpression studies.

This study represents the first comprehensive analysis of *ELF5* expression at the isoform level, using RNA-sequencing (RNA-seq) data from The Cancer Genome Atlas (TCGA) for 6757 normal tissue and cancer samples. The functional effects of ELF5 isoform expression in breast cancer were also investigated using inducible cell line models and a 116-gene quantitative polymerase chain reaction (qPCR) panel, leading to unique insights into the transcriptional functions of ELF5 and in particular the role of the PNT domain.

## Methods

### RNA-sequencing analysis

RNA-Seq version 2 data for initial primary tumors and solid tissue normal samples (where  $n \geq 3$ ) were downloaded from TCGA data portal (<https://tcga-data.nci.nih.gov/tcga/>) [29–43], with institutional human research ethics committee exemption. Samples with available RNA-Seq version 2 data (August 2013 for breast and April 2014 for all other cancer types) were included. The RNA-Seq version 2 TCGA pipeline for preprocessing of publicly available data used MapSplice [44] for alignment and RSEM [45] for quantitation. Non-normalized gene and isoform data were downloaded from TCGA as RSEM expected (“raw”) counts, unadjusted for transcript length, and scaled estimates, adjusted for transcript length. Scaled estimates were multiplied by  $10^6$  to obtain transcripts per million (TPM) values. Normalized gene and isoform data were downloaded from TCGA as quantile normalized RSEM expected counts (unadjusted for transcript length), with the upper quartile set at 1000 for gene data and 300 for isoform data.

A summary of all TCGA samples used in the analysis is shown in Table 1. For breast cancer samples, PAM50

**Table 1** Summary of all TCGA RNA sequencing samples used in analysis

Tissue	Cancer type	TCGA acronym	Normal samples <sup>a</sup>	Cancer samples
Bladder	Bladder urothelial carcinoma	BLCA	19	241
Breast	Breast invasive carcinoma	BRCA	59 <sup>b</sup>	515
	Luminal A			229
	Luminal B			126
	HER2			57
	Basal-like			96
	Normal-like			7
Cervix	Cervical squamous cell carcinoma and endocervical adenocarcinoma	CESC	3	185
Colon	Colon adenocarcinoma	COAD	41	261
Head/neck (including mouth and throat)	Head and neck squamous cell carcinoma	HNSC	43	497
Kidney	Chromiophobe	KICH	25	66
	Clear cell carcinoma	KIRC	72	518
	Papillary cell carcinoma	KIRP	30	172
Liver	Hepatocellular carcinoma	LIHC	50	191
Lung	Lung adenocarcinoma	LUAD	56	488
	Lung squamous cell carcinoma	LUSC	50	490
Pancreas	Pancreatic adenocarcinoma	PAAD	3	85
Prostate	Prostate adenocarcinoma	PRAD	50	297
Rectum	Rectum adenocarcinoma	READ	9	91
Thyroid	Thyroid carcinoma	THCA	39	498
Uterus	Uterine corpus endometrial carcinoma	UCEC	24	158
	Uterine carcinosarcoma	UCS	NA <sup>c</sup>	57
Adrenal gland	Adrenocortical carcinoma	ACC	NA	79
Hematological	Diffuse large B-cell lymphoma	DLBC	NA	28
	Acute myeloid leukemia	AML		173
Brain	Glioblastoma multiforme	GBM	NA	156
	Lower grade glioma	LGG		463
Ovary	Ovarian serous cystadenocarcinoma	OV	NA	262
Skin	Cutaneous melanoma	SKCM	NA	82
Bone/connective tissue/soft tissue	Sarcoma	SARC	NA	103

TCGA The Cancer Genome Atlas

<sup>a</sup>Normal samples included where  $n \geq 3$ <sup>b</sup>65 samples included in differential expression analysis<sup>c</sup>Uterine corpus endometrioid carcinoma normal samples used as normal uterine samples for differential expression analysis

(Predication Analysis of Microarrays 50-gene classifier) status was used to generate a subtyped cohort of 515 patients and 59 matched normal samples [29, 46]. Six additional normal samples, matching to tumors in the initial cohort, were included in differential expression analyses.

Limma voom [47] was used for differential expression analysis of gene-level RNA-seq data, with inputs as non-normalized gene data (RSEM expected counts). Filtering was applied to remove genes with low expression, keeping genes with counts  $>1$  in at least  $n$  samples (where  $n$  = number of samples in smallest group of

replicates). The trimmed mean of M-values normalization method [48] was applied, followed by differential expression analysis using Limma voom. All fold change (FC) and false discovery rate (FDR) values reported were generated by Limma voom analyses. Venn diagrams were created using online software (<http://bioinformatics.psb.ugent.be/webtools/Venn/>), and clustered heat maps were created using the R package gplots [49]. As a comparison, differential expression analysis was also carried out using edgeR [50–54] (see Additional file 1: Methods).

### Stable cell line generation

*ELF5* isoforms 1, 2, and 3 were tagged with C-terminal V5 (and short linker sequence), cloned into the pHUSH-ProEx vector [55], and used as a retrovirus. T47D-EcoR and MDA-MB-231-EcoR cells stably expressing ecotropic receptor were infected with pHUSH-*ELF5* retrovirus and selected using puromycin. To generate clonal cell lines, stable cell line pools were plated at low density in 96-well plates.

### Cell lines and treatments

All cell lines were obtained from the American Type Culture Collection (Manassas, VA, USA) and were maintained in RPMI medium supplemented with insulin and 10 % tetracycline-free fetal bovine serum (Clontech Laboratories, Mountain View, CA, USA). Puromycin was added at a concentration of 1 µg/ml. Doxycycline (Dox) was added at a concentration of 0.1 µg/ml daily to induce protein expression.

### Cell number assay

Cell number was quantified using a spectrophotometric assay. Cells were incubated with 16 % trichloroacetic acid and stained with 10 % Diff-Quik II solution (Lab Aids, Narrabeen, Australia). 10 % acetic acid was added to dried plates, and 100 µl of solution from each well was added to a 96-well plate, which was read at 595 nm. Absorbance readings were transformed to natural logarithms, and values from three wells (single experiment) were averaged for each time point. The minus Dox and plus Dox slopes for each cell line were compared using linear regression analysis.

### Western blot analysis

Protein was prepared in NuPAGE Sample Buffer and Reducing Agent (Life Technologies, Carlsbad, CA, USA) using 10 µg (estrogen-related blots), 65 µg (V5 blot, T47D-*ELF5*-isoform 2-V5) or 25 µg (V5 blots, all other lines) per lane. Samples were separated on precast 15-well 4–12 % Bis-Tris (estrogen-related blots) or 10-well 10 % Bis-Tris (V5 blots) polyacrylamide gels (Life Technologies), transferred to polyvinylidene fluoride membrane, blocked in 5 % skim milk, and incubated overnight at 4 °C in primary antibody. Secondary horseradish peroxidase-conjugated antibody was added 1:2000 in 5 % skim milk (anti-mouse, NA931V, anti-rabbit, NA934V; GE Healthcare Life Sciences, Little Chalfont, UK). Proteins were detected using enhanced chemiluminescence solution (Western Lightning Plus; PerkinElmer, Waltham, MA, USA) and x-ray film (Fujifilm, Tokyo, Japan). Primary antibodies used were anti-V5 (sc-58052, 1:500–1:1000; Santa Cruz Biotechnology, Santa Cruz, CA, USA), anti-transducin-like enhancer of split 1 (anti-TLE1) (ab183742, 1:1000; Abcam, Cambridge, UK), anti-ERα (sc-8005, 1:1000;

Santa Cruz Biotechnology), anti-Forkhead box A1 (anti-FOXA1) (sc-101058, 1:1000; Santa Cruz Biotechnology), and anti-β-actin (AC-15, 1:20,000; Sigma-Aldrich, St. Louis, MO, USA).

### Transient retroviral infection

*ELF5* isoform 3 was tagged with C-terminal hemagglutinin (HA), cloned into the pQCXIH vector (Clontech) and used as a retrovirus. MDA-MB-231-EcoR-pHUSH-*ELF5*-isoform 2-V5 Clone 7 cells were infected with *ELF5*-isoform 3-HA/empty vector retrovirus diluted 1:4. No pQCXIH selection pressure was applied.

### Immunofluorescence

Cells were infected with pQCXIH retrovirus in eight-well Lab-Tek II chamber slides (Thermo Scientific, Waltham, MA, USA) and allowed to recover for 24 h. Dox /vehicle treatment (lasting 24 h) was then commenced. Cells were fixed with 4 % paraformaldehyde diluted in PHEM buffer (60 mM piperazine-*N,N'*-bis(2-ethanesulfonic acid) (PIPES), 25 mM 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid (HEPES), 1 mM ethylene glycol tetraacetic acid (EGTA), 2 mM MgCl<sub>2</sub>, pH 6.9), permeabilized with 0.5 % Triton X-100, blocked with 10 % donkey serum/PHEM solution, and incubated overnight at 4 °C in primary antibody. Secondary antibodies were added at 1:200, and coverslips were applied using Duolink In Situ Mounting Medium with 4',6-diamidino-2-phenylindole (DAPI) (Olink Bioscience, Uppsala, Sweden). Imaging was performed on a Leica DM5500 microscope (Leica Microsystems, Wetzlar, Germany). Antibodies (in 10 % donkey serum/PHEM solution): anti-V5 (sc-58052, 1:200; Santa Cruz Biotechnology), anti-HA (3724, 1:800; Cell Signaling Technology, Danvers, MA, USA), and donkey anti-mouse Alexa Fluor 647 and donkey anti-rabbit Alexa Fluor 555 conjugates (1:200; Molecular Probes/Thermo Fisher Scientific, Eugene, OR, USA).

### Quantitative PCR

RNA was extracted using the RNeasy Mini Kit with DNase treatment (Qiagen, Valencia, CA, USA) and quantified using the NanoDrop spectrophotometer (NanoDrop Products, Wilmington, DE, USA). Complementary DNA (cDNA) was made using the High-Capacity cDNA Reverse Transcription Kit (Life Technologies) with ribonuclease inhibitor (Promega, Madison, WI, USA). All qPCRs were run on an ABI 7900 qPCR machine (Applied Biosystems, Foster City, CA, USA), using standard TaqMan cycling conditions or Roche Universal Probe Library (UPL) protocol with two or three technical replicates per sample (see also Additional file 1).

For the clonal cell line time-course qPCR (Fig. 6f), 0.5 µg of RNA per 20 µl of cDNA reaction and *ELF5* (Hs01063022\_m1) and glyceraldehyde 3-phosphate



dehydrogenase (4236317E) assays were used. For the 116-gene panel, cell lines were treated for 48 h with Dox or vehicle. cDNA reactions were scaled to 100  $\mu$ l and 2.5  $\mu$ g RNA. Roche UPL assays were designed using the online Roche ProbeFinder software. All assays are detailed in Additional file 2.

Results were analyzed using SDS 2.4 (Life Technologies) and qbase+ software (Biogazelle, Gent, Belgium) [56]. Paired *t* tests were used to calculate *p* values, comparing -Dox and +Dox samples (three or four pairs per cell line group). Correction for multiple comparisons was performed using the Benjamini-Hochberg procedure, setting the FDR at 0.10 [57].

## Results

### *ELF5* isoforms are differentially expressed in normal tissues

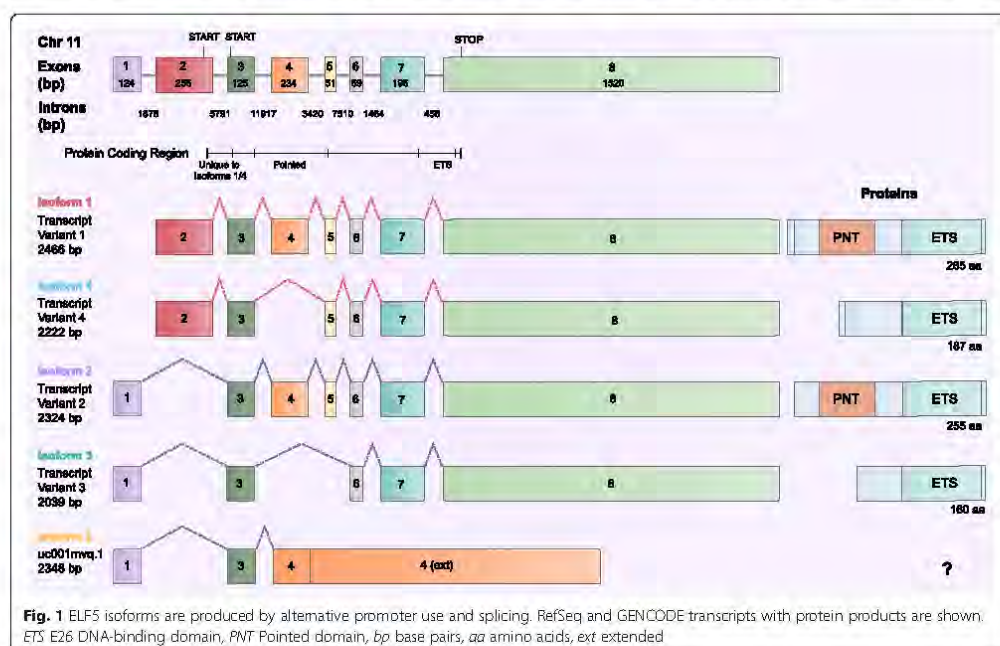
There are four *ELF5* transcript variants in the National Center for Biotechnology Information RefSeq database [58], predicted to produce four unique proteins (Fig. 1). The two full-length transcripts (isoforms 1 and 2) use alternative promoters, resulting in unique first exons and proteins that differ by only ten N-terminal amino acids. Two additional transcripts (isoforms 3 and 4) are produced by splicing of exons 4 ( $\pm 5$ ) from each of the full-length transcripts, producing proteins that lack the PNT domain but retain the ETS domain. An additional transcript (isoform 5), described by the GENCODE Consortium [59],

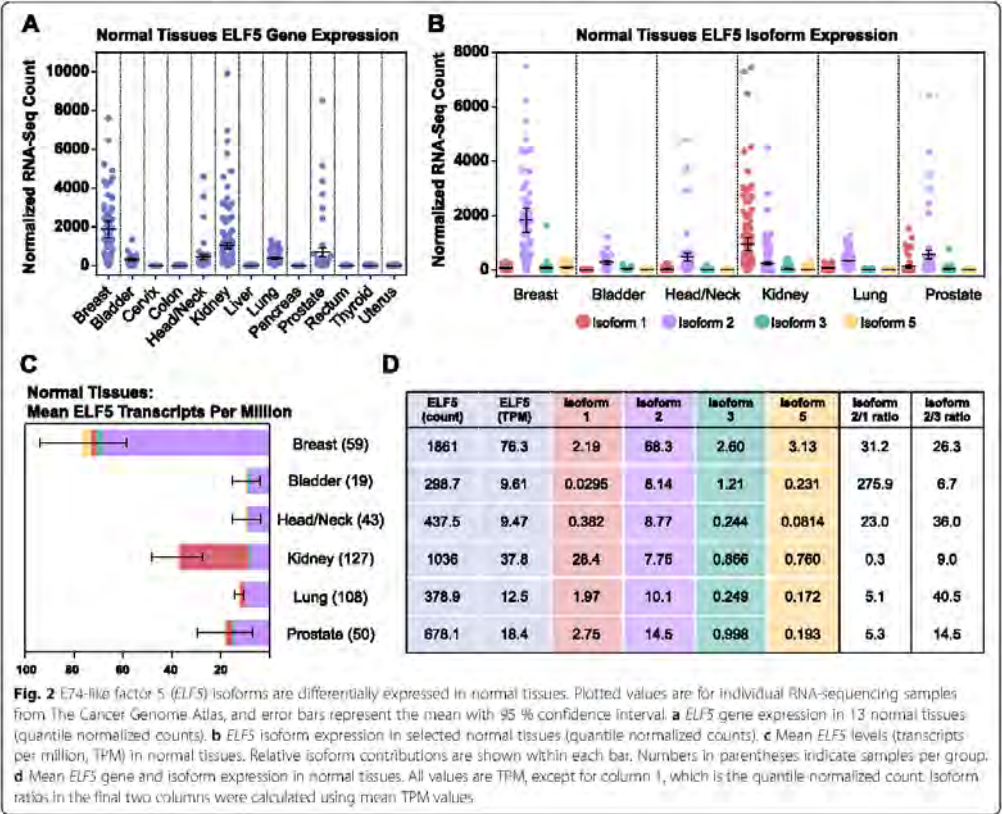
is a variant of isoform 2 terminating at an extended exon 4. This type of intronic extension ("bleeding exon") is often associated with incompletely processed transcripts [60], and it is unclear whether this transcript produces a protein product (which would lack the ETS domain).

RNA-seq data from TCGA were analyzed to quantify and compare *ELF5* isoforms in normal and cancer tissues [29–43]. A summary of all TCGA samples analyzed is shown in Table 1. TCGA preprocessed data include *ELF5* isoforms 1, 2, and 3 as annotated by RefSeq, as well as isoform 5. Due to the reference annotation used by TCGA, there are no data for *ELF5* isoform 4. The transcripts and protein products are summarized in Fig. 1, and a cross-database comparison is shown in Additional file 1: Figure S1.

*ELF5* expression was highest in epithelial tissues such as the breast, kidney, lung, prostate, and bladder (Fig. 2a). The breast was one of the highest *ELF5*-expressing tissues in the body. Isoform 1 and 2 expression was highly tissue-specific (Fig. 2b), indicating alternative promoter use in different tissues.

Data in Fig. 2a and b were quantile-normalized by the TCGA pipeline, allowing comparison of abundance of a particular transcript (such as total *ELF5*) between samples. However, longer transcripts will generate more sequencing reads, making quantitative comparison of transcripts of different lengths problematic. To overcome this, the proportional measure TPM may be used. TPM is an





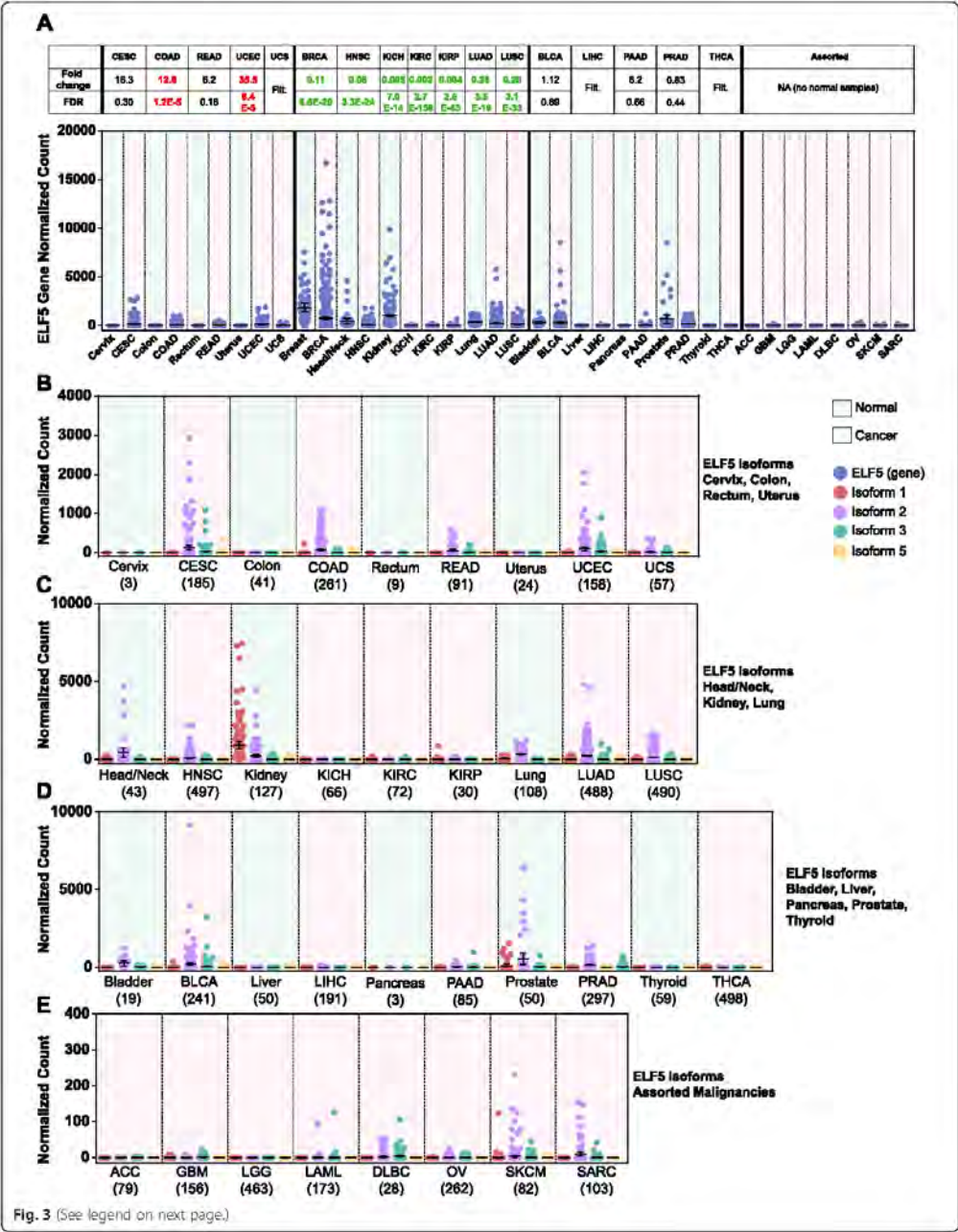
example of a within-sample normalization method, and it should be noted that values are not technically comparable between samples, particularly when the composition of the total mRNA pool may be quite different (for example, when comparing different tissues). For this reason, data are shown for both quantile-normalized (between-samples-normalized) (Fig. 2b) and TPM-normalized (within-sample-normalized) (Additional file 1: Figure S2b). As the lengths of *ELF5* transcripts are not widely different, ranging from 2039 to 2466 base pairs, the data plots are in fact similar.

Since TPM is a proportional measure, the relative abundances of transcripts of different lengths within samples can be compared. The mean TPM values for *ELF5* isoforms are shown in Fig. 2c and d. Breast, bladder, head/neck, lung, and prostate all expressed isoform 2 as their main transcript (median percentage 82.1–95.2 %) (Additional file 1: Figure S2a), while the kidney expressed mainly isoform 1 (median 91.8 %). All tissues examined expressed, on average, more full-length isoform 2 than the shorter isoform 3.

***ELF5* expression is significantly altered in cancer**

In malignancy, *ELF5* expression was significantly altered compared with normal tissues, as shown by Limma voom differential gene expression analysis (Fig. 3a). In the cervix, colon, rectum, and uterus, cancer was associated with an increase in *ELF5* level, driven mainly by an increase in isoform 2 and, to a lesser extent, isoform 3 (Fig. 3b). Conversely, there was almost complete suppression of *ELF5* expression in three kidney carcinoma subtypes. *ELF5* expression was also significantly decreased in head and neck, lung, and prostate cancer (Fig. 3c). In both lung carcinoma subtypes, there was a large variation in *ELF5* levels, suggesting possible molecular subtype-specific expression patterns, similar to the breast. *ELF5* expression was largely unchanged (or filtered from analysis due to low expression) in the tissues shown in Fig. 3d. The cancer types shown in Fig. 3e exhibited very low levels of *ELF5* expression but had no normal tissue samples available for comparison. Analysis of additional RNA-seq normal tissue datasets (Genotype-Tissue





(See figure on previous page.)

**Fig. 3** E74-like factor-5 (*ELF5*) expression is significantly altered in cancer. The Cancer Genome Atlas (TCGA) RNA-sequencing (RNA-seq) data for 25 cancer types (pink background) are shown, with normal tissue comparisons (green background) where available. Plotted values are for individual TCGA RNA-seq samples, and error bars represent the mean with 95 % confidence interval. TCGA cancer acronyms are used (see Table 1). **a** *ELF5* gene expression (normalized counts) for 25 cancers with normal tissue comparisons where available. Fold changes and false discovery rates (FDRs) from Limma voom analysis are shown, with green values in bold indicating significant downregulation and red values in bold significant upregulation compared with normal (FDR < 0.05). *Filter* indicates gene filtered from Limma voom analysis due to low expression. **b** *ELF5* isoform expression in normal and cancer samples (with *ELF5* gene upregulation in cancer). **c** *ELF5* isoform expression in normal and cancer samples (with *ELF5* gene downregulation in cancer). **d** *ELF5* isoform expression in normal and cancer samples (unchanged or filtered *ELF5*). **e** *ELF5* isoform expression in cancer samples without available normal samples (normal samples ≤ 2). Note smaller scale on y-axis

Expression Project and Illumina Human BodyMap) confirmed that the normal adrenal gland, brain, leukocytes/whole blood, lymph node, ovary, and skeletal muscle all had very low or absent *ELF5* expression (Additional file 1: Figure S3a and b). Skin was the only exception from this group of tissues demonstrating moderate *ELF5* expression consistent with previous studies of differentiated keratinocytes [6].

Differential expression analysis was also carried out using edgeR. Overall, the results from Limma voom and edgeR were similar. The edgeR FC and FDR values are presented in Additional file 1: Figure S4a for comparison.

#### *ELF5* expression is altered in breast cancer in a subtype-specific manner

Comprehensive analysis of RNA-seq incorporating molecular subtype was undertaken for 515 breast cancer patients. In the luminal A, luminal B, and HER2 subtypes, *ELF5* was significantly downregulated (fold change 0.02–0.13 compared to normal), while in the basal subtype there was a strong trend for increased *ELF5* expression (1.96-fold compared with normal, FDR 0.053 in Limma voom analysis, 1.99-fold compared with normal, FDR 0.0008 in edgeR analysis) (Fig. 4a and Additional file 1: Figure S4b). There was no clear relationship between *ELF5* expression and American Joint Committee on Cancer stage (Additional file 1: Figure S5).

This analysis was extended to the isoform level by examining the contribution to total *ELF5* (based on mean TPM) for each isoform (Fig. 4b). Normal-like samples were excluded due to low sample numbers. The main isoform expressed in all breast cancer subtypes was isoform 2. In the luminal A, luminal B, and HER2 subtypes, all *ELF5* isoforms were decreased in cancer compared with normal (Fig. 4c). Conversely, in the basal subtype, three of four isoforms were upregulated, with isoform 3 having a relatively larger fold change.

The percentage contributions of each isoform to total *ELF5* were also analyzed (Fig. 4d and e). The normal breast showed a tight range of expression, while in cancer, particularly for isoforms 2 and 3, this was broadened (Fig. 4d). The high variability in isoform 3 percentage values in the cancer samples led to an increased mean percentage in all subtypes. Median values demonstrated a smaller, although still

increased, isoform 3 percentage in cancer, while the median isoform 2 percentage remained fairly constant across normal and cancer samples.

Within this cohort, 65 patients had matched tumor and normal samples that could be directly compared (Fig. 4f and Additional file 1: Figure S4c). The luminal A, luminal B, and HER2 groups showed a highly significant decrease in *ELF5* level in both the Limma and edgeR analyses. In the basal subgroup, there was an upward but variable trend.

#### Expression of other ETS family members is also altered in breast cancer, with the basal subtype having a distinct ETS expression profile

The same cohort of patients was used to examine expression of other members of the ETS transcription factor family. RNA-seq data showed that a large number of ETS factors were expressed in the normal breast. Average TPM values (which take into account transcript length) for ETS factors in the normal breast ranged from 0.02 to 117.7. Several ETS factors had very low expression (<2 TPM), including *FEV*, *SPIC*, *ETV2*, *ETV3L*, and *SPIB*. The most highly expressed ETS factors in the normal breast were *EHF*, *ELF3*, *SPDEF*, and *ELF5* (Additional file 1: Figure S6).

ETS factor expression was significantly altered in breast cancer, as shown by Limma voom differential expression analysis. In the first (unpaired) analysis, samples from each molecular subtype, excluding normal-like, were compared with the common set of 65 normal breast samples, allowing analysis of larger sample sets. In the second (paired) analysis, normal and subtyped tumor samples from the same patient were compared, allowing for more rigorously matched comparisons but limited by smaller sample numbers. ETS factors with low expression (three to five per subtype) were filtered from the analysis.

Of the 25 ETS factors included in the unpaired analysis, 24 were significantly altered in at least 1 subtype, with 14 common to all subtypes (Fig. 5a). Within these, 13 were altered in the same direction (5 up and 8 down in the tumor compared with normal), while *SPDEF* was oppositely regulated in basal compared with other subtypes. In the paired analysis, 21 ETS factors were significantly altered in at least 1 subtype, with 3 ETS factors common to all subtypes (*SPDEF*, *ERG*, and *ETS2*) and an additional 8 common to 3 of 4 subtypes (Fig. 5b). *ELF5*

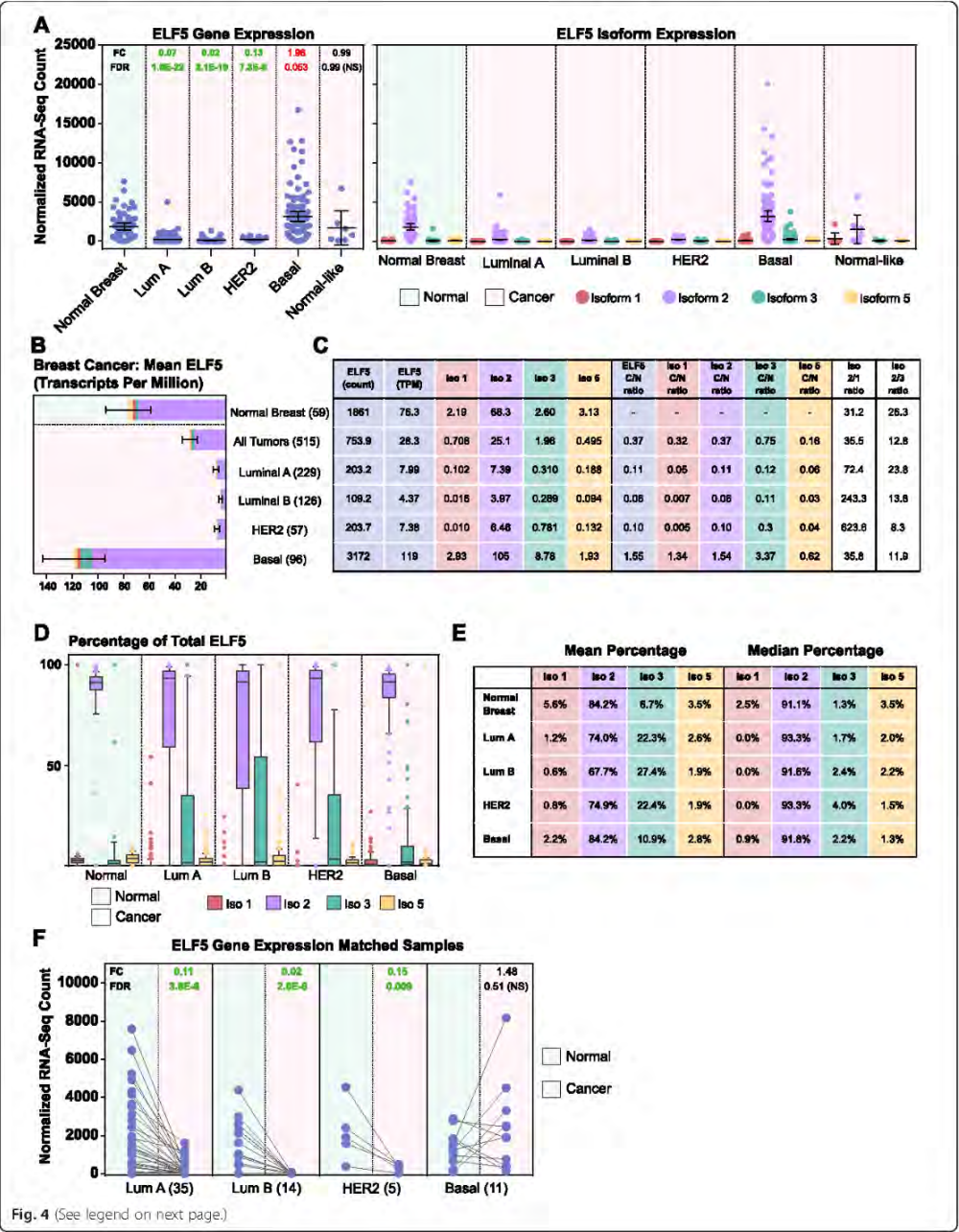


Fig. 4 (See legend on next page.)



(See figure on previous page.)

**Fig. 4** E74-like factor 5 (*ELF5*) expression is altered in breast cancer in a subtype-specific manner. **a** *ELF5* gene (left) and isoform (right) expression (quantile-normalized counts) for normal breast and breast cancer subtypes. Plotted values are for The Cancer Genome Atlas RNA-sequencing (RNA-seq) samples, and error bars represent the mean with 95 % confidence interval. Fold change (FC) and false discovery rate (FDR) from Limma voom analysis are shown for *ELF5* gene data, with green values in bold indicating a significant downregulation and red values in bold a significant upregulation compared with normal (FDR < 0.05). Nonbold green or red values indicate FDR of 0.05–0.10. **b** Mean *ELF5* levels in transcripts per million (TPM) in normal breast and breast cancer, excluding normal-like, with 95 % confidence interval. Relative isoform contributions shown within each bar. Numbers in parentheses indicate samples per group. **c** Mean *ELF5* expression values at the gene and isoform levels (columns 1–6), isoform fold changes in cancer compared with normal (columns 7–11), and isoform ratios (columns 12 and 13). All values are TPM, except for column 1, which is the quantile-normalized count. Ratios were calculated using mean TPM values. **d** Box-and-whisker plot representing isoform percentage of total *ELF5* in normal breast and cancer. Box 25–75th percentile, horizontal line median, error bars 10th–90th percentile, circles outliers. **e** Mean (left) and median (right) isoform percentage values for normal breast and cancer. **f** *ELF5* levels (quantile-normalized count) for patients with matched normal and cancer samples, categorized according to tumor molecular subtype. Six extra matched normal samples were included, for a total of 65 pairs. Plotted values represent individual samples, with samples from the same patient connected with a line. FC and FDR from paired Limma voom analysis are shown, with green values indicating a significant downregulation compared with normal (FDR < 0.05). Numbers in parentheses indicate sample pairs per group

was the most downregulated ETS family member by fold change in the luminal A, luminal B, and HER2 subtypes in both unpaired and paired analyses.

Compared with other subtypes, the basal group showed a number of unique ETS factor expression changes. To further explore this, the Limma *t* statistics for all ETS family members (tumor compared with normal) were plotted on a clustered heat map (Fig. 5c, unpaired, and Fig. 5d, paired). The basal subtype showed a distinct expression profile and clustered separately from the other subtypes in both paired and unpaired analyses, highlighting the potential for the ETS transcription factor family to exert a unique transcriptional influence in this subtype. Similar results were obtained with unpaired and paired edgeR analyses (Additional file 1: Figure S7).

Several ETS family members with significant changes in expression were selected to visualize the results of the breast cancer differential expression analyses. The normalized counts for *ERG* (downregulated), *ETV7* (upregulated), and *SPDEF* (differentially regulated) are shown in Fig. 5e. Direct comparison of matched normal and tumor samples is shown in Fig. 5f. Interestingly, *SPDEF* showed the inverse expression pattern of *ELF5*. The normalized counts for the entire ETS factor family, with the results of the Limma voom and edgeR differential expression analysis, are shown in Additional file 1: Figure S8.

#### Alterations in cell line *ELF5* isoform levels result in a similar phenotype, characterized by decreased cell number, decreased estrogen-related proteins, and nuclear localization

TCGA data showed an increased diversity of *ELF5* isoform expression in cancer compared with the normal breast; therefore, the expression levels and effects of *ELF5* isoform expression were examined in vitro to determine if this was of functional consequence.

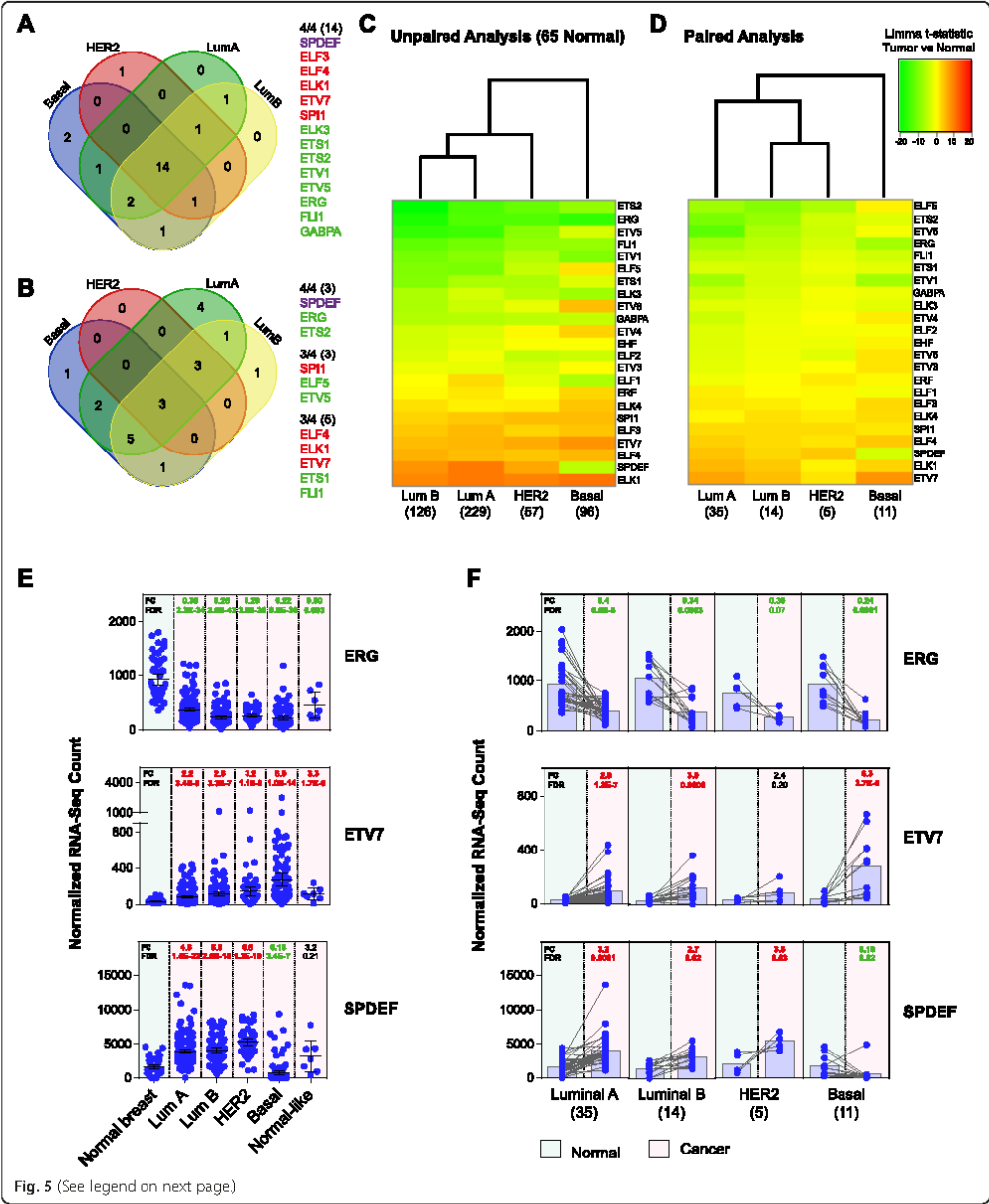
*ELF5* expression in a panel of breast cancer cell lines was analyzed by qPCR and Western blotting (Additional file 1: Figure S9a and d). Three cell lines (T47D, BT474,

and HCC1187) expressed high levels of *ELF5* protein (Additional file 1: Figure S9d), with the size of the main band consistent with isoform 2. A possible band representing isoform 3 was seen in the HCC1187 cell line; however, interpretation was difficult due to high background.

Clonal cell lines were constructed with a Dox-inducible expression vector containing a single *ELF5* isoform, tagged with C-terminal V5. The luminal cell line T47D (ER+/PR+/HER2-) was chosen to examine the effect of isoforms in the context of relatively high endogenous *ELF5* expression, testing the hypothesis that isoforms lacking the PNT domain might exert a dominant-negative effect on full-length isoform function. A second claudin-low cell line, MDA-MB-231 (ER-/PR-/HER2-), was chosen as it expresses no endogenous *ELF5*, allowing the effects of each isoform to be determined in the absence of potential competitive isoform interactions.

Over a 5-day time course, induced expression of isoforms 1, 2, and 3 all resulted in a significantly decreased growth rate in T47D cells, with no change in the empty vector control (Fig. 6a). Representative light microscopic images for T47D lines (Fig. 6b) demonstrate decreased cell number and increased detached cells (additional images shown in Additional file 1: Figure S9e and f). A similar but less pronounced decrease in growth rate was also seen with induction of isoform 2 and isoform 3 in the MDA-MB-231 lines (Fig. 6c). It has previously been shown that the mechanisms underlying this phenotype for *ELF5* isoform 2 include G<sub>1</sub> arrest, increased apoptosis, and reduced adhesion proteins [25].

In the T47D lines, each isoform caused a decrease in ERα protein and pioneer factors FOXA1 and TLE1, required for ER–chromatin interactions [61, 62] (Fig. 6d). The effects on FOXA1 and TLE1 were also seen in the MDA-MB-231 lines, in the absence of detectable ERα (Fig. 6e). Dox-inducible *ELF5* mRNA expression was shown by qPCR (day 5) (Fig. 6f). V5 antibody Western blot analysis confirmed *ELF5*-V5 protein expression and also illustrated the size difference between isoforms 2 and 3 (Fig. 6g).



(See figure on previous page.)

**Fig. 5** Expression of other E26 transforming sequence (ETS) family members is also altered in breast cancer, with the basal subtype having a distinct ETS expression profile. The Cancer Genome Atlas RNA-sequencing (RNA-Seq) Limma voom differential expression analysis data for ETS family members. **a** Venn diagram showing number of ETS family members significantly altered in breast cancer subtypes compared with normal (false discovery rate (FDR) < 0.05). All subtypes were compared with a common set of 65 normal samples (unpaired analysis). Genes altered in all four subtypes are listed (red = upregulation, green = downregulation, purple = differentially regulated in basal subtype compared with other subtypes). **b** Venn diagram showing number of ETS family members significantly altered in breast cancer subtypes compared with normal (FDR < 0.05), using paired normal and tumor samples from the same patient. Genes altered in at least three of four subtypes are listed, with color-coding as above. **c** Clustered heat map of ETS factor Limma voom *t* statistic, comparing tumor samples to the common set of 65 normal samples. Legend is shown next to **(d)**. Rows are sorted by luminal B values (smallest to largest), and columns are sorted according to clustering. Numbers in parentheses are samples per group. **d** Clustered heat map of Limma voom *t* statistic, comparing paired normal and tumor samples, with sorting as above. Numbers in parentheses are sample pairs per group. **e** Expression of *ERG*, *ETV7*, and *SPDEF* for normal breast (green background) and breast cancer subtypes (pink background). Plotted values are for individual samples (normalized counts), and error bars represent the mean with 95 % confidence interval. Fold change (FC) and FDR from unpaired Limma voom differential expression analysis are shown, with green indicating a significant downregulation and red a significant upregulation compared with normal (FDR < 0.05). **f** *ERG*, *ETV7*, and *SPDEF* levels for a 65 patients with matched normal and cancer samples. FC and FDR from paired Limma voom differential expression analysis are shown, with color-coding as above (FDR < 0.05). Numbers in parentheses are sample pairs per group

Immunofluorescence was performed to determine the subcellular location of ELF5 isoforms when expressed in isolation and when coexpressed. MDA-MB-231 cells with Dox -inducible ELF5-isoform 2-V5 expression were used, with transient retroviral infection of an ELF5-isoform 3-HA vector. This allowed manipulation of isoform 2 and isoform 3 levels within the same cell. Figure 6h (top row) shows MDA-MB-231-ELF5-isoform 2-V5 cells treated with Dox to induce expression, as well as transient infection of a control pQCXIH vector. There was strong nuclear V5 staining and no HA staining. In row 2, cells were treated with Dox to induce ELF5-isoform 2-V5 and also infected with isoform 3-HA. Both isoform 2 (V5) and isoform 3 (HA) localized to the nucleus, and there was no cytoplasmic redistribution seen in the cells that expressed both isoform 2 and isoform 3 (indicated by arrows), an effect that has been reported previously for ETS1 isoforms [63].

#### ELF5 isoforms have a similar transcriptional effect in T47D and MDA-MB-231 cell lines

A panel of 116 genes was examined by qPCR to compare the transcriptional effects of ELF5 isoforms. Previously published microarrays and ELF5/V5 chromatin immunoprecipitation with massively parallel DNA sequencing [25] were used to identify genes and pathways regulated by ELF5 isoform 2 in luminal cell lines. The assays are described in Additional file 2, with an outline of the experimental workflow shown in Additional file 1: Figure S10.

The pHUSH clonal cell lines were selected on the basis of similar qPCR levels of ELF5 isoform induction. Figure 7a shows the ELF5 level with Dox relative to the without Dox control for each individual cell line. To compare baseline (without Dox) variability, values were also normalized to the lowest ELF5 value (Fig. 7b). Baseline variability was minimal in the T47D lines; however, expression ranged from 1.0- to 2.3 in the MDA-MB-231 isoform 3 lines and from 4.7 (clone 6) to 28.0 (clone 1)

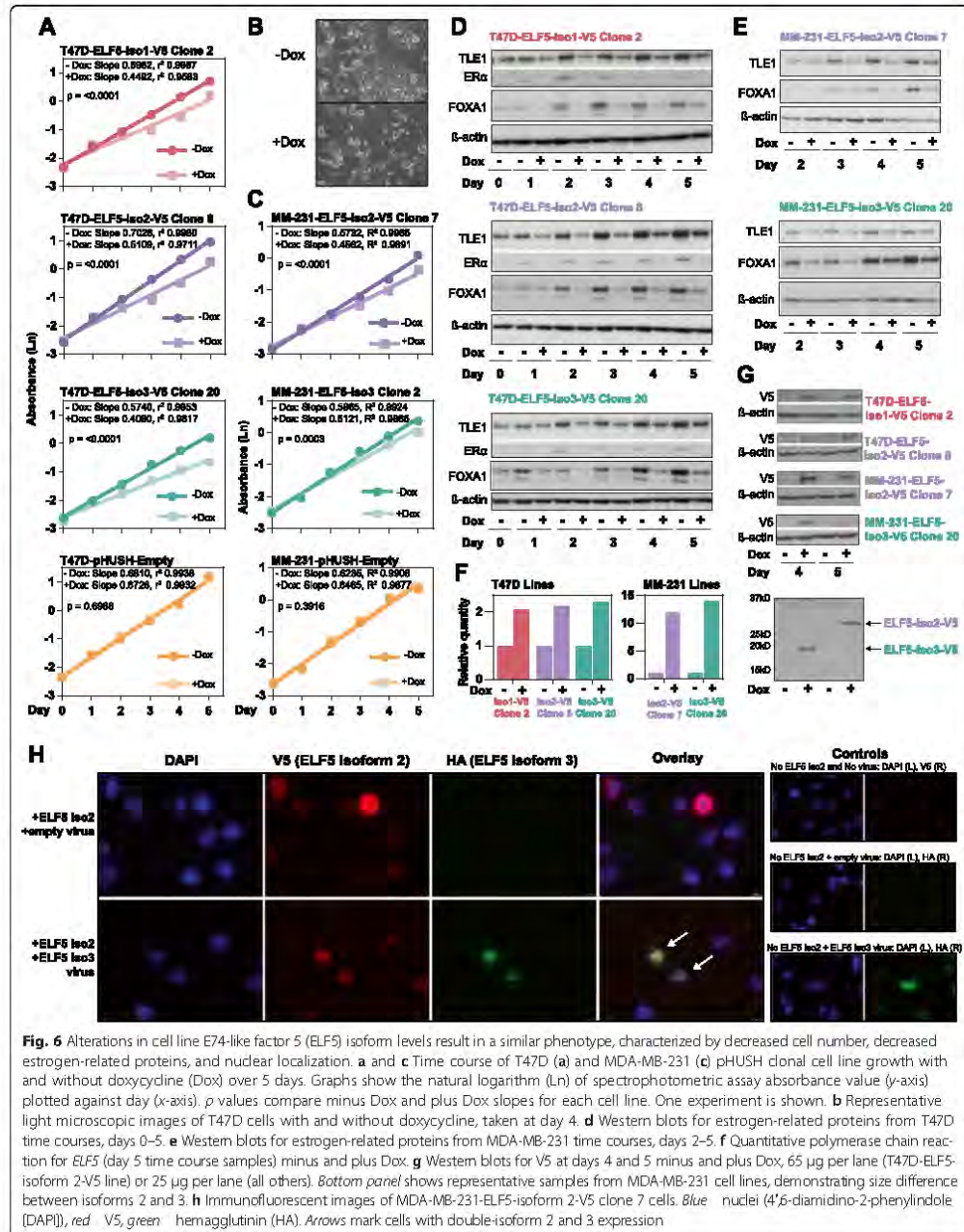
in the isoform 2 lines. This variation is most likely due to slight “leakiness” of the pHUSH vector, leading to low-level ELF5 expression (undetectable by V5 Western blotting) in the absence of Dox.

T47D and MDA-MB-231 clonal cell lines were treated with Dox or vehicle for 48 h to induce ELF5 isoform expression. Initially, two clones per parental cell line were used. A selection of 27 genes was then repeated in 1 or 2 further clones, giving a total of 3 or 4 clonal lines (biological replicates) per parental line (Additional file 1: Table S1). The heat maps in Fig. 7 show the  $\log_{10}$  FC for each gene when ELF5 isoform expression is induced (+dox) compared with baseline (-dox).

Overall, the pattern of change was fairly similar, regardless of which ELF5 isoform was expressed. The genes with the strongest absolute FC (>3 in any T47D line or >2 in any MDA-MB-231 line) showed a particularly consistent pattern of change (Fig. 7c). Expression changes were greater in the T47D than in the MDA-MB-231 cell lines.

Genes were also analyzed in functional categories (Fig. 7d). Apoptosis-related genes showed consistent changes corresponding to an increase in apoptosis, such as upregulation of apoptosis-promoting genes, including *DDIT3*, *PUMA*, *NOXA*, *TP53*, and various caspases, as well as downregulation of apoptosis-inhibiting genes such as *BCLX* and *BCL2*. The changes in cell cycle genes were weaker, although still generally consistent, with upregulation of cell cycle inhibitors such as *RB1CC1* and *TP53* and downregulation of cell cycle-promoting genes such as cyclins D1, B1, A2, and E2 and associated kinases *CDK1/2*. However, the pattern of change was not entirely congruent with inhibition of the cell cycle, with upregulation of the cyclin D-associated *CDK6* and downregulation of the cell cycle inhibitor *CDKN2C* (p18). Changes in mRNA expression for key genes associated with estrogen action, such as *ESR1*, *FOXAI*, *GATA3*, and *GREB1*, were relatively small and variable (Fig. 7d), in contrast to results at the protein level,





**Fig. 6** Alterations in cell line E74-like factor 5 (ELF5) isoform levels result in a similar phenotype, characterized by decreased cell number, decreased estrogen-related proteins, and nuclear localization. **a** and **c** Time course of T47D (**a**) and MDA-MB-231 (**c**) pHUSH clonal cell line growth with and without doxycycline (Dox) over 5 days. Graphs show the natural logarithm (Ln) of spectrophotometric assay absorbance value (y-axis) plotted against day (x-axis).  $p$  values compare minus Dox and plus Dox slopes for each cell line. One experiment is shown. **b** Representative light microscopic images of T47D cells with and without doxycycline, taken at day 4. **d** Western blots for estrogen-related proteins from T47D time courses, days 0–5. **e** Western blots for estrogen-related proteins from MDA-MB-231 time courses, days 2–5. **f** Quantitative polymerase chain reaction for *ELF5* (day 5 time course samples) minus and plus Dox. **g** Western blots for V5 at days 4 and 5 minus and plus Dox, 65 µg per lane (T47D-ELF5-isoform 2-V5 line) or 25 µg per lane (all others). **h** Western blots shows representative samples from MDA-MB-231 cell lines, demonstrating size difference between isoforms 2 and 3. **i** Immunofluorescent images of MDA-MB-231-ELF5-isoform 2-V5 clone 7 cells. Blue = nuclei (4',6-diamidino-2-phenylindole [DAPI]), red = V5, green = hemagglutinin (HA). Arrows mark cells with double-isoform 2 and 3 expression

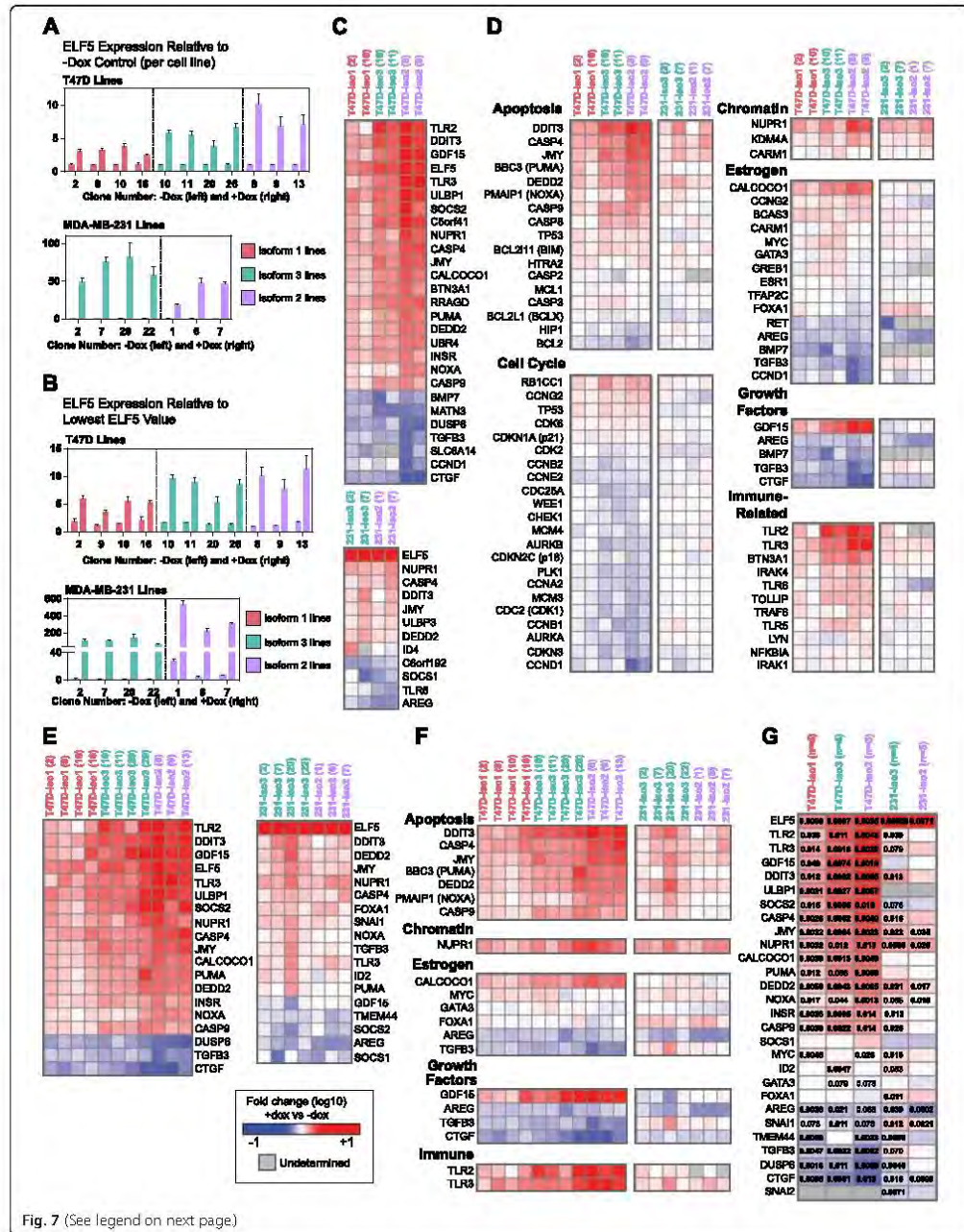


Fig. 7 (See legend on next page.)

(See figure on previous page.)

**Fig. 7** E74-like factor 5 (ELF5) isoforms have a similar transcriptional effect in T47D and MDA-MB-231 cell lines. **a** ELF5 expression measured by quantitative polymerase chain reaction (qPCR) at 48 h for T47D clonal cell lines (top) and MDA-MB-231 clonal cell lines (bottom). Assay detects all ELF5 isoforms. Values are the mean calibrated normalized relative quantity (CNQ) with standard error. Results relative to the minus doxycycline (–Dox) control (set at 1) for each cell line. **b** ELF5 expression measured by qPCR at 48 h for T47D clonal cell lines (top) and MDA-MB-231 clonal cell lines (bottom) used in the qPCR panel. Assay detects all ELF5 isoforms. Values are the mean CNQ with standard error. Results relative to the sample with the lowest ELF5 value (set at 1), which is T47D-ELF5-isoform 2-VS clone 8 (T47D lines) and MDA-MB-231-ELF5-isoform 3-VS clone 22 (MDA-MB-231 lines). **c** Heat map showing genes (from 116-gene qPCR panel) with absolute fold change >3 (any T47D line) or >2 (any MDA-MB-231 line). Two clonal cell lines were tested per group. All heat maps use the legend shown in (e) and represent the log<sub>10</sub> fold change (capped at –1 and +1) of the plus Dox quantity compared with the minus Dox quantity as measured by qPCR. Gray indicates gene was not detectable by qPCR in minus and/or plus Dox samples. **d** Functional categorization of selected genes from 116-gene qPCR panel. Some genes are represented more than once due to multiple functions. **e** Heat map showing genes (from 27-gene qPCR panel) with absolute fold change >3 (any T47D line) or >2 (any MDA-MB-231 line). Results shown for three or four clonal lines per group. **f** Functional categorization of selected genes from 27-gene qPCR panel. **g** Heat map representing the mean log<sub>10</sub> fold change per group for all genes in the 27-gene panel, as well as ELF5. Significant *p* values are shown where false discovery rate (FDR) is <0.10. Some *p* values (nonhold) are >0.05, although FDR for these values is <0.10. Nonsignificant *p* values (FDR >0.10) are not shown

which showed robust downregulation of ESR1 and FOXA1 with all ELF5 isoforms.

The results were substantiated using 1 or 2 further clones per parental cell line and 27 genes from the original panel (Fig. 7e and f). The average FC for each parental cell line group (consisting of three or four clonal cell lines) was calculated, and this is shown in the heat map in Fig. 7g with corresponding significant *p* values (FDR <0.10). Although the pattern of change was generally consistent, there were some interesting differences. First, *FOXA1* expression in the T47D lines exhibited a mostly downward trend, although there were no statistically significant changes. Conversely, in the MDA-MB-231 lines, *FOXA1* mRNA increased (significant only in the isoform 2 group); again, this is in contrast to the protein results shown for the MDA-MB-231 lines in Fig. 7g. Second, there was only one case in the T47D lines (and none in the MDA-MB-231 lines) in which a gene was altered in statistically significant opposite directions by different ELF5 isoforms. This gene, *GATA3*, was upregulated by isoform 3 and downregulated by isoform 2, although the changes were relatively small. In fact, 20 of 27 genes in the T47D lines showed a statistically significant change in the same direction with each of the 3 isoforms, pointing toward the overall consistency of the transcriptional effect of ELF5 isoforms.

## Discussion

This study is the first detailed analysis of ELF5 isoform expression and function, extending previous ELF5 Northern blot analysis, immunohistochemistry, and microarray studies [5, 6, 16, 25] to the isoform level using 6757 sequenced normal and cancer samples. The kidney appears to be unique in being the only tissue examined to express isoform 1 as its dominant isoform, expanding on the initial Northern blot analysis-based descriptions of ELF5 isoforms [6]. In breast cancer, ELF5 alterations were subtype-specific, with the basal subtype demonstrating unique ELF5 isoform expression changes. Despite

differences in protein domains, the in vitro phenotypic and transcriptional effects of increased ELF5 isoform expression were similar. This suggests that ELF5 action is regulated in various tissues by tissue-specific alternative promoter use rather than by differences in the transcriptional activity of the isoforms.

In cancer, ELF5 expression is frequently altered. The kidney, one of the highest ELF5-expressing tissues, showed a dramatic decrease in ELF5 level in cancer. ELF5 has been characterized as a tumor suppressor in the kidney and bladder [19, 20], and this may restrict kidney carcinomas to non-ELF5-expressing cells of origin. In other tissues, cancer was associated with an aberrant increase in ELF5 expression, as seen in the cervix, colon, rectum, and uterus. This may indicate an oncogenic role for ELF5 in these tissues or broader genomic deregulation, such as DNA hypomethylation, a hallmark of the cancer genome [64]. The mechanisms regulating ELF5 in different tissues and in cancer have not been widely studied; however, in the early embryo and the developing mammary gland, ELF5 regulation of lineage specification is associated with promoter methylation status [65, 66]. Increased ELF5 promoter methylation has also been demonstrated in bladder carcinoma [19]. These studies establish DNA methylation as an important epigenetic mechanism regulating ELF5 expression, with possible aberrant methylation in cancer.

The normal human breast expresses relatively high levels of ELF5, with subtype-specific alterations in cancer. High ELF5 has been shown to maintain the ER– basal phenotype, paralleling the normal developmental role of specification of the ER– alveolar lineage [25]. In all breast cancer subtypes, there was a broader distribution of ELF5 isoform expression. Increased variability of isoform distribution (“transcriptome instability”) is a known phenomenon and is proposed as a molecular hallmark of cancer [67, 68]. A recent study identified 244 cancer-associated isoform “switches” involving consistent changes in the most abundant isoform [69]. An ELF5 isoform switch has not been



identified in breast cancer, in keeping with the present study, which showed an inconsistent pattern of isoform expression variation. Although not consistently identified, this does not mean that ELF5 isoform switches do not play an important role in the subset of patients in which they occur.

Other ETS transcription factors have also been shown to be important in breast cancer. Extension of RNA-seq analysis to the entire ETS family revealed a number of cancer-associated expression changes. The ETS family as a whole has previously been studied in breast cancer at the qPCR level in mouse models [70] and human cell lines [71], although the present study is the first, to our knowledge, to include examination of the expression of the entire human ETS family in both the normal breast and subtyped breast cancer samples using RNA-seq data. The normal human breast expressed a diverse range of ETS factors. Compared with the normal breast, the basal-like subtype showed a distinct pattern of ETS factor expression changes, with several ETS factors changing in the opposite direction in basal compared with other subtypes. *ELF5* and *SPDEF* were the most striking examples of this phenomenon. *SPDEF* is also a luminal epithelial lineage-specific transcription factor in the breast and has been shown to promote the survival of ER+ breast cancer cells [72]. The inverse relationship seen between these two transcription factors in breast cancer is intriguing and may well have a parallel during normal mammary development.

Finally, the phenotypic and transcriptional effects of isoforms 1, 2, and 3 were found to be similar in inducible cell line models. This was unexpected, as the PNT domain in murine *ELF5* has previously been shown to have strong transactivation activity [12]. In many proteins, SAM and/or PNT domains act as protein-protein interaction modules, an important mechanism of biological specificity for ETS factors, which often bind only weakly to DNA in the absence of binding partners or posttranslational modifications [3, 12]. The importance of the PNT domain is also shown by other ETS family members in which removal of the PNT domain significantly alters protein function. The endogenous ETS1 isoform p27, for example, lacks the PNT and transactivation domains and negatively regulates full-length ETS1 by competing for DNA-binding sites and promoting its translocation from the nucleus to the cytoplasm [63]. Although this splicing event is similar to those that occur to produce *ELF5* isoforms 3 and 4, it appears that *ELF5* isoform 3 can alter gene transcription in a very similar way to the full-length isoforms. In addition, there was no subcellular relocation of full-length isoform 2 seen when isoform 3 was coexpressed. Interestingly, however, while exogenous *ELF5* localized to the nucleus in this study, cytoplasmic *ELF5* staining is seen in some human breast cancer

samples and is a predictor of outcome [73]. This indicates that endogenous *ELF5* can localize to the cytoplasm and that this has functional significance in breast cancer. A potential nuclear export sequence exists in the ETS domain of *ELF5* (amino acids 165–174) similar to one identified in *ELF3* [74, 75]. It is possible that cytoplasmic relocation of *ELF5* is mediated by the relative amounts of isoforms but that this effect is not recapitulated by exogenous expression, particularly in the context of MDA-MB-231 cells, which do not normally express *ELF5* and therefore may be lacking essential protein binding partners. Given the importance of context in the function of ETS factors, it is possible that the differential effects of *ELF5* isoforms may also require a stimulus (for example, growth factors) or challenge (for example, estrogen deprivation) in order to become apparent, an avenue that was not explored in this study.

## Conclusions

This study has characterized the expression pattern and functions of *ELF5* at the isoform level, demonstrating significantly altered expression in cancer. Alterations in *ELF5* isoform expression in cancer may drive abnormal cell fate decisions, suggesting that *ELF5*, like other ETS factors, may be a significant contributor to tumorigenesis. While further studies are needed to clarify the mechanisms that regulate differential *ELF5* isoform expression and to fully elucidate the role of the PNT domain, understanding expression and function at the isoform level is a vital first step in realizing the potential of transcription factors such as *ELF5* as prognostic markers or therapeutic targets in cancer.

## Additional files

**Additional file 1:** Methods for additional figures, extended methods for qPCR experiments, Table S1 describing all clonal cell lines, additional Figures S1–S10 including legends as described in the main text. (PDF 11918 kb)

**Additional file 2:** Excel spreadsheet with details of all qPCR assays (Roche UPL assays worksheet 1 and TaqMan assays worksheet 2). (XLSX 25 kb)

## Abbreviations

cDNA: complementary DNA; CNRQ: Calibrated Normalized Relative Quantity; Dox: doxycycline; *ELF5*: E74-like factor 5; ER: estrogen receptor- $\alpha$ ; ETS: E26 transforming sequence; FC: fold change; FDR: false discovery rate; FOXA1: Forkhead box A1; HA: hemagglutinin; HER2: Erb-b2 receptor tyrosine kinase 2; PAM50: Prediction Analysis of Microarrays 50-gene classifier; PNT: Pointed domain; PR: progesterone receptor; qPCR: quantitative polymerase chain reaction; RNA-seq: RNA sequencing; SAM: sterile alpha motif domain; TCGA: The Cancer Genome Atlas; TGF- $\beta$ : transforming growth factor- $\beta$ ; TLE1: transducing-like enhancer of split 1; TPM: transcripts per million, a proportional measure of abundance correcting for transcript length; UPL: Roche Universal Probe Library.

## Competing interests

The authors declare that they have no competing interests.

### Authors' contributions

CLP performed RNA-sequencing data analyses, in vitro functional studies, and drafting of the manuscript. DLR assisted with bioinformatics and drafting of the manuscript. HJL, DGO, and SRO assisted with in vitro experiments and revision of the manuscript. CJO conceived of the study and its design and participated in the drafting of the manuscript. All authors read and approved the final manuscript.

### Acknowledgments

This work was supported by the Australian Postgraduate Award (University of New South Wales), grants from the National Health and Medical Research Council Australia (project 1047149 and fellowship 1043400), Banque Nationale de Paris-Paribas Australia and New Zealand, R.T. Hall Trust, and the National Breast Cancer Foundation (fellowships ECF-13-08 and PF-12-06 and award NC-12-24).

### Author details

<sup>1</sup>Cancer Division, Garvan Institute of Medical Research/The Kinghorn Cancer Centre, Sydney, NSW 2010, Australia. <sup>2</sup>Babraham Institute, Cambridge CB22 3AT, UK.

Received: 6 October 2015 Accepted: 16 December 2015

Published online: 07 January 2016

### References

- Oikawa T, Yamada T. Molecular biology of the Ets family of transcription factors. *Gene*. 2003;303:11–34.
- Graves BJ, Petersen JM. Specificity within the Ets family of transcription factors. *Adv Cancer Res*. 1998;75:1–55.
- Li R, Pei H, Watson DK. Regulation of Ets function by protein-protein interactions. *Oncogene*. 2000;19(55):6514–23.
- Kar A, Gutierrez-Hartmann A. Molecular mechanisms of ETS transcription factor-mediated tumorigenesis. *Crit Rev Biochem Mol Biol*. 2013;48(6):522–43.
- Zhou J, Ng AY, Tymms MJ, Jemlin LS, Seth AK, Thomas RS, et al. A novel transcription factor, ELF5, belongs to the ELF subfamily of ETS genes and maps to human chromosome 11p13-15, a region subject to LOH and rearrangement in human carcinoma cell lines. *Oncogene*. 1998;17(21):2719–32.
- Oertgen P, Kas K, Dube A, Gu X, Grall F, Thamrongsak U, et al. Characterization of ESE-2, a novel ESE-1-related Ets transcription factor that is restricted to glandular epithelium and differentiated keratinocytes. *J Biol Chem*. 1999;274(41):29439–52.
- Schultz J, Milpetz F, Bork P, Ponting CP. SMART, a simple modular architecture research tool identification of signaling domains. *Proc Natl Acad Sci U S A*. 1998;95(11):5857–64.
- Kim CA, Phillips ML, Kim W, Gingery M, Tran HH, Robinson MA, et al. Polymerization of the SAM domain of TEL in leukemogenesis and transcriptional repression. *EMBO J*. 2001;20(15):4173–82.
- Seidel JJ, Graves BJ. An ERK2 docking site in the Pointed domain distinguishes a subset of ETS transcription factors. *Genes Dev*. 2002;16(1):127–37.
- Green JB, Gardner CD, Wharton RP, Aggarwal AK. RNA recognition via the SAM domain of Smaug. *Mol Cell*. 2003;11(6):1537–48.
- Barrera FN, Poveda JA, González-Ros JM, Neira JL. Binding of the C-terminal sterile  $\alpha$  motif (SAM) domain of human p73 to lipid membranes. *J Biol Chem*. 2003;278(47):46878–85.
- Choi YS, Sinha S. Determination of the consensus DNA-binding sequence and a transcriptional activation domain for ESE-2. *Biochem J*. 2006;398(3):497–507.
- Donnison M, Beaton A, Davey HW, Broadhurst R, L'Huillier P, Pfeffer PL. Loss of the extraembryonic ectoderm in Elf5 mutants leads to defects in embryonic patterning. *Development*. 2005;132(10):2299–308.
- Metzger DE, Stahlman MT, Shannon JM. Misexpression of ELF5 disrupts lung branching and inhibits epithelial differentiation. *Dev Biol*. 2008;320(1):149–60.
- Oakes SR, Naylor MJ, Asselin-Labat ML, Blazek KD, Gardiner-Garden M, Hilton HN, et al. The Ets transcription factor ELF5 specifies mammary alveolar cell fate. *Genes Dev*. 2008;22(5):581–6.
- Lapinskas EJ, Palmer J, Ricardo S, Hertzog PJ, Hammacher A, Pritchard MA. A major site of expression of the ets transcription factor Elf5 is epithelia of exocrine glands. *Histochem Cell Biol*. 2004;122(6):521–6.
- Yao B, Zhao J, Li Y, Li H, Hu Z, Pan P, et al. Elf5 inhibits TGF- $\beta$ -driven epithelial-mesenchymal transition in prostate cancer by repressing SMAD3 activation. *Prostate*. 2015;75(8):872–82.
- Xie BX, Zhang H, Wang J, Pang B, Wu RQ, Qian XL, et al. Analysis of differentially expressed genes in LNCaP prostate cancer progression model. *J Androl*. 2011;32(2):170–82.
- Wu B, Cao X, Liang X, Zhang X, Zhang W, Sun G, et al. Epigenetic regulation of Elf5 is associated with epithelial-mesenchymal transition in urothelial cancer. *PLoS One*. 2015;10(1), e0117510.
- Lapinskas EJ, Svobodova S, Davis ID, Cebon J, Hertzog PJ, Pritchard MA. The Ets transcription factor ELF5 functions as a tumor suppressor in the kidney. *Twin Res Hum Genet*. 2011;14(4):316–22.
- Risinger JI, Maxwell GL, Chandramouli GV, Jazayeri A, Aprelikova O, Patterson T, et al. Microarray analysis reveals distinct gene expression profiles among different histologic types of endometrial cancer. *Cancer Res*. 2003;63(1):6–11.
- Panagopoulos I, Gorunova L, Davidson B, Heim S. Novel TNS3-MAP3K3 and ZFPM2-ELF5 fusion genes identified by RNA sequencing in multicystic mesothelioma with t(7;17)(p12;q23) and t(8;11)(q23;p13). *Cancer Lett*. 2015;357(2):502–9.
- Tomlins SA, Rhodes DR, Perner S, Dhanasekaran SM, Mehra R, Sun XW, et al. Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science*. 2005;310(5748):644–8.
- Perou CM, Sorlie T, Eisen MB, van de Rijn N, Jeffrey SS, Rees CA, et al. Molecular portraits of human breast tumours. *Nature*. 2000;406(6797):747–52.
- Kalyuga M, Gallego-Ortega D, Lee HJ, Roden DL, Cowley MJ, Calkon CE, et al. ELF5 suppresses estrogen sensitivity and underpins the acquisition of antiestrogen resistance in luminal breast cancer. *PLoS Biol*. 2012;10(12), e1001461.
- Chakrabarti R, Hwang J, Andres Blanco M, Wei Y, Lukacisin M, Romano RA, et al. Elf5 inhibits the epithelial-mesenchymal transition in mammary gland development and breast cancer metastasis by transcriptionally repressing Snail2. *Nat Cell Biol*. 2012;14(11):1212–22.
- Pal S, Gupta R, Davuluri RV. Alternative transcription and alternative splicing in cancer. *Pharmacol Ther*. 2012;136(3):283–94.
- Bates DO, Cui TG, Doughty JM, Winkler M, Sugiono M, Shields JD, et al. VEGF165b, an inhibitory splice variant of vascular endothelial growth factor, is down-regulated in renal cell carcinoma. *Cancer Res*. 2002;62(14):4123–31.
- Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature*. 2012;490(7418):61–70.
- Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*. 2008;455(7216):1061–8. A published corrigendum appears in *Nature*. 2013;494(7438):506.
- Brennan CW, Verhaak RG, McKenna A, Campos B, Nourbakhsh H, Salama SR, et al. The somatic genomic landscape of glioblastoma. *Cell*. 2013;155(2):462–77.
- Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. *Nature*. 2011;474(7353):609–15.
- Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers. *Nature*. 2012;489(7417):519–25.
- Cancer Genome Atlas Research Network. Comprehensive molecular profiling of lung adenocarcinoma. *Nature*. 2014;511(7511):543–50.
- Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. *Nature*. 2012;487(7407):330–7.
- Cancer Genome Atlas Research Network. Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature*. 2013;499(7456):43–9.
- Cancer Genome Atlas Research Network. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *N Engl J Med*. 2013;368(22):2059–74.
- Cancer Genome Atlas Research Network. Integrated genomic characterization of endometrial carcinoma. *Nature*. 2013;497(7447):67–73. A published erratum appears in *Nature*. 2013;500(7461):242.
- Cancer Genome Atlas Research Network. Comprehensive molecular characterization of urothelial bladder carcinoma. *Nature*. 2014;507(7492):315–22.
- Cancer Genome Atlas Research Network. Integrated genomic characterization of papillary thyroid carcinoma. *Cell*. 2014;159(3):676–90.
- Davis CF, Ricketts CJ, Wang M, Yang L, Cherniack AD, Shen H, et al. The somatic genomic landscape of chromophobe renal cell carcinoma. *Cancer Cell*. 2014;26(3):319–30.
- Cancer Genome Atlas Network. Genomic classification of cutaneous melanoma. *Cell*. 2015;161(7):1681–96.
- Cancer Genome Atlas Network. Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature*. 2015;517(7536):576–82.
- Wang K, Singh D, Zeng Z, Coleman SJ, Huang Y, Savich GL, et al. MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Res*. 2010;38(18):e178.

45. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*. 2011;12:323.
46. Parker JS, Mullins M, Cheung MC, Leung S, Vouduc D, Vickery T, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. *J Clin Oncol*. 2009;27(8):1160–7.
47. Law CW, Chen Y, Shi W, Smyth GK. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol*. 2014;15(2):R29.
48. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol*. 2010;11(3):R25.
49. Warnes GR, Bolker B, Bonebakker L, Gentleman R, Huber W, Liaw A, et al. gplots: Various R Programming Tools for Plotting Data. R package version 2.17.0. <https://cran.r-project.org/web/packages/gplots/index.html>. Accessed 22 Dec 2015.
50. Robinson MD, Smyth GK. Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics*. 2007;23(21):2881–7.
51. Robinson MD, Smyth GK. Small-sample estimation of negative binomial dispersion, with applications to SAGE data. *Biostatistics*. 2008;9(2):321–32.
52. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*. 2010;26(1):139–40.
53. McCarthy DJ, Chen Y, Smyth GK. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res*. 2012;40(10):4288–97.
54. Zhou X, Lindsay H, Robinson MD. Robustly detecting differential expression in RNA sequencing data using observation weights. *Nucleic Acids Res*. 2014;42(11):e91.
55. Gray DC, Hofflich KP, Peng L, Gu Z, Gogineni A, Murray LJ, et al. pHUSH: a single vector system for conditional gene expression. *BMC Biotechnol*. 2007;7:61.
56. Hellebrand J, Mortier G, De Paeppe A, Speleman F, Vandesompele J. qBase: relative quantification framework and software for management and automated analysis of real-time quantitative PCR data. *Genome Biol*. 2007;8(2):R19.
57. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Series B Stat Methodol*. 1995;57(1):289–300.
58. Pruitt K, Brown G, Tatusova T, Maglott J. The reference sequence (RefSeq) database. In: McEntyre J, Ostell J, editors. *The NCBI handbook* [Internet]. Bethesda, MD: National Library of Medicine, National Center for Biotechnology Information; 2002. <http://www.ncbi.nlm.nih.gov/books/NBK21091/> [last update 6 Apr 2012; accessed 22 Dec 2015].
59. Hanow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, et al. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res*. 2012;22(9):1760–74.
60. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The human genome browser at UCSC. *Genome Res*. 2002;12(6):996–1006.
61. Hurtado A, Holmes KA, Ross-Innes CS, Schmidt D, Carroll JS. FOXA1 is a key determinant of estrogen receptor function and endocrine response. *Nat Genet*. 2011;43(1):27–33.
62. Holmes KA, Hurtado A, Brown GD, Launchbury R, Ross-Innes CS, Hadfield J, et al. Transducin-like enhancer protein 1 mediates estrogen receptor binding and transcriptional activity in breast cancer cells. *Proc Natl Acad Sci U S A*. 2012;109(8):2748–53.
63. Laitem C, Leprieux G, Chouli S, Begue A, Monte D, Larimont D, et al. Ets-1 p27: a novel Ets-1 isoform with dominant-negative effects on the transcriptional properties and the subcellular localization of Ets-1 p51. *Oncogene*. 2009;28(20):2087–99.
64. Gama-Sosa MA, Slagel VA, Trewn RW, Oxenhandler R, Kuo KC, Gehrke CW, et al. The 5-methylcytosine content of DNA from human tumors. *Nucleic Acids Res*. 1983;11(19):6883–94.
65. Lee HJ, Hinshelwood RA, Bouras T, Gallego-Ortega D, Valdés-Mora F, Blazek K, et al. Lineage specific methylation of the *Elf5* promoter in mammary epithelial cells. *Stem Cells*. 2011;29(10):1611–9.
66. Ng RK, Dean W, Dawson C, Lucifero D, Madeja Z, Reik W, et al. Epigenetic restriction of embryonic cell lineage fate by methylation of *Elf5*. *Nat Cell Biol*. 2008;10(11):1280–90.
67. Venables JP, Klinck R, Koh C, Gervais-Bird J, Bramard A, Intel L, et al. Cancer-associated regulation of alternative splicing. *Nat Struct Mol Biol*. 2009;16(6):670–6.
68. Sveen A, Johannessen B, Teixeira MR, Lothe RA, Skotheim RI. Transcriptome instability as a molecular pan-cancer characteristic of carcinomas. *BMC Genomics*. 2014;15:672.
69. Sebestyén E, Zawisza M, Eyrás E. Detection of recurrent alternative splicing switches in tumor samples reveals novel signatures of cancer. *Nucleic Acids Res*. 2015;43(3):1345–56.
70. Galang CK, Müller WJ, Foos G, Oshima RG, Hauser CA. Changes in the expression of many Ets family transcription factors and of potential target genes in normal mammary tissue and tumors. *J Biol Chem*. 2004;279(12):11281–92.
71. He J, Pan Y, Hu J, Albarracín C, Wu Y, Dai JL. Profile of Ets gene expression in human breast carcinoma. *Cancer Biol Ther*. 2007;6(1):76–82.
72. Buchwalter G, Hickey MM, Cromer A, Seifors LM, Gunawardane RN, Frisman J, et al. PDEF promotes luminal differentiation and acts as a survival factor for ER-positive breast cancer cells. *Cancer Cell*. 2013;23(6):753–67.
73. Gallego-Ortega D, Ledger A, Roden D, Liaw AM, Magenau A, Kikhtyak Z, et al. ELF5 drives lung metastasis in luminal breast cancer through recruitment of Gr1<sup>+</sup>CD11b<sup>+</sup> myeloid-derived suppressor cells. *PLoS Biol*. 2013;13(12):e1002330. doi:10.1371/journal.pbio.1002330.
74. Prescott JD, Koto KS, Singh M, Gutierrez-Hartmann A. The ETS transcription factor ESE-1 transforms MCF-12A human mammary epithelial cells via a novel cytoplasmic mechanism. *Mol Cell Biol*. 2004;24(12):5548–64.
75. Prescott JD, Pocaschutt JM, Tentler JJ, Walker DM, Gutierrez-Hartmann A. Mapping of ESE-1 subdomains required to initiate mammary epithelial cell transformation via a cytoplasmic mechanism. *Mol Cancer*. 2011;10:103.

Submit your next manuscript to BioMed Central and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at  
[www.biomedcentral.com/submit](http://www.biomedcentral.com/submit)

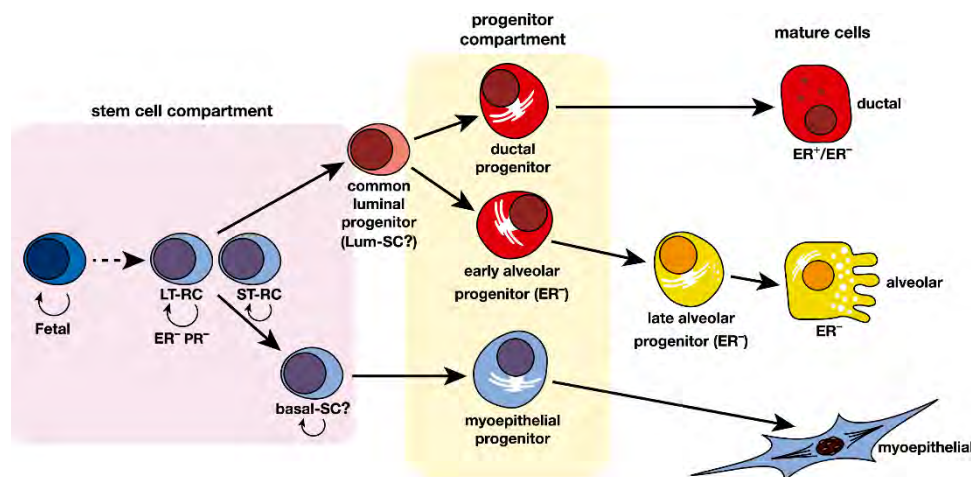


## Chapter 4: Genome-wide Studies of ELF5 Action

### Introduction

#### ELF5 regulates cell fate in normal development and breast cancer

ELF5 is a master regulator of cell fate in the mammary gland epithelium. Although many details of the mammary epithelial cell differentiation hierarchy remain controversial, most models describe the differentiation of the mammary stem cell into the luminal progenitor cell and subsequently into two main types of mature luminal cells, the ductal (or hormone-sensory) cells and alveolar secretory cells (Figure 4.1) (reviewed in Visvader and Stingl, 2014). This differentiation hierarchy, underpinned by the long-lived mammary stem cell, facilitates the extensive cycles of proliferation, differentiation and involution that occur during every pregnancy.



**Figure 4.1: Model of the mammary epithelial hierarchy**

*Figure adapted from (Visvader and Stingl, 2014)*

Hypothetical model of the mammary epithelial hierarchy, comprising stem cells, progenitor cells and mature cells. The stem cell compartment in the adult mammary gland contains multipotent long-term and short-term repopulating cells (LT-RCs and ST-RCs), which lack hormone receptor expression. Stem cells give rise to both luminal and myoepithelial progenitors, with the existence of unipotent luminal- and basal-specific stem cells still debated. The balance between ELF5 and ER transcriptional activities is proposed to guide differentiation of the luminal progenitor cell into the alveolar (ER-negative, ELF5-high) and ductal (ER-positive, ELF5-low) lineages.

A defining characteristic of the mammary stem cell is the ability to regenerate an entire functional mammary gland in murine transplant experiments (Shackleton *et al.*, 2006; Stingl *et al.*, 2006). ELF5 expression is suppressed in the mammary stem cell due to methylation of its promoter. However, as the cell progresses towards an epithelial cell fate, the ELF5 promoter is partially demethylated and the expression of ELF5 rises (Gallego-Ortega *et al.*, 2013). It is hypothesised that the increase in ELF5 levels may in fact be what drives this first cell fate decision in the mammary gland, as experimental manipulation of ELF5 levels results in alterations in the stem cell compartment consistent with a role for ELF5 in mammary stem cell differentiation (Chakrabarti *et al.*, 2012b; Lee *et al.*, 2013).

The luminal progenitor cell is poised between the hormone-sensory and secretory lineages and the balance between ELF5 and ER transcriptional activities is hypothesised to determine the ultimate cell fate decision. A further rise in ELF5 expression, for example, in response to hormonal cues such as progesterone-induced paracrine signalling, promotes the differentiation of the luminal progenitor cell into a mature ER/PR-negative secretory cell. In the absence of these cues, the ER transcriptional network prevails, and the cell differentiates into a hormone receptor-positive cell with low ELF5 expression. In the normal mammary gland, therefore, ELF5 promotes the development of the luminal progenitor cell into an ER/PR-negative cell population that further expands during pregnancy to form the milk-producing cells of the alveoli (Lee *et al.*, 2013; Oakes *et al.*, 2008).

The luminal progenitor cell is also the likely cell-of-origin for most types of breast cancer (reviewed in Gross *et al.*, 2016). Importantly, the developmental roles of ELF5, and the interplay with ER, appear to be mirrored in breast cancer. ELF5 has been shown to suppress the luminal (ER-positive) phenotype and to promote the growth of oestrogen-insensitive breast cancer cells, including basal-like cells and endocrine-resistant luminal cells (Kalyuga *et al.*, 2012). A model for ELF5 action in breast cancer proposes that the balance between ELF5 and ER activity in the evolving cancer cell determines breast cancer subtype (basal-like or luminal), paralleling normal development. Subsequent changes in ELF5 expression may prompt additional cell fate alterations, such as an ELF5-driven shift towards an oestrogen-insensitive phenotype in a luminal breast cancer that results in resistance to ER-targeting therapies such as tamoxifen (Gallego-Ortega *et al.*, 2013). In both the luminal and basal-like subtypes, increased ELF5 expression suppresses the mesenchymal (Claudin-low) subtype, consistent with the developmental role in promoting mesenchymal-to-epithelial



differentiation (Chakrabarti *et al.*, 2012a; Kalyuga *et al.*, 2012).

The mechanisms by which ELF5 reduces the oestrogen sensitivity of luminal breast cancer cells are not completely understood. One mechanism is the ELF5-mediated decrease in the expression of numerous ER-activated genes, thereby opposing the regulatory effects of the ER/FOXA1 network. ELF5 also down-regulates ER itself, as well as the ER pioneer factor FOXA1. Using chromatin immunoprecipitation followed by sequencing (ChIP-seq), ELF5 was shown to bind to the FOXA1 promoter in the luminal cell line T47D, suggesting that ELF5 directly represses FOXA1 expression (Kalyuga *et al.*, 2012).

This chapter investigates the effects of increased ELF5 expression in a luminal breast cancer context using next-generation sequencing technology, with the ultimate aim of uncovering the molecular mechanisms by which ELF5 promotes an oestrogen-insensitive cell fate. The ER-positive cell line MCF7 was selected for study as this cell line normally expresses a low level of ELF5, and this is increased in the context of *in vitro* tamoxifen resistance (Kalyuga *et al.*, 2012). The MCF7 cell line therefore provides a suitable model for examining the effects of elevated ELF5 expression on modulation of oestrogen sensitivity.

The effects of the most commonly-expressed ELF5 isoform (Isoform 2, as shown in Chapter 3) were examined at several levels. Firstly, the ELF5 transcriptome was investigated using next-generation RNA-sequencing, which currently represents the gold standard for high-sensitivity global transcriptomic studies (Wang *et al.*, 2014). This was combined with ELF5 chromatin immunoprecipitation followed by DNA sequencing (ChIP-seq) to detect the genomic locations of ELF5 binding. Integration of RNA-seq and ChIP-seq data then allowed the identification of a subset of genes likely to be directly regulated by ELF5 binding. Finally, FOXA1 ChIP-seq was performed in the context of low and high ELF5 expression to determine if ELF5 could influence the genomic patterns of FOXA1 binding, a novel potential mechanism for modulating the endocrine response.

## Results

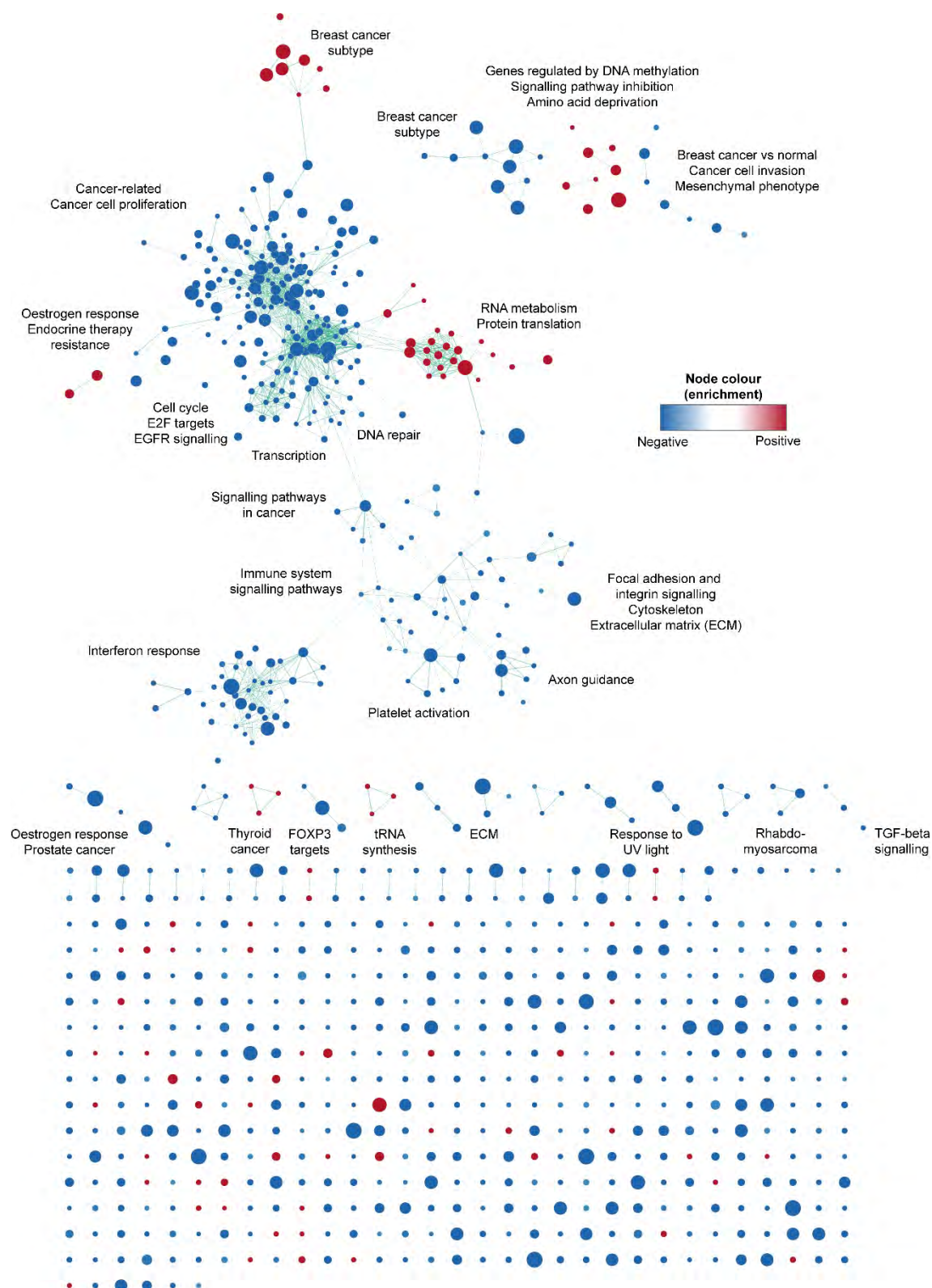
### Identification of ELF5-regulated genes using RNA sequencing

MCF7-pHUSH-ELF5-Isoform2-V5 (MCF7-ELF5-V5) cells were treated with doxycycline (or vehicle) to induce ELF5-Isoform2-V5 (ELF5-V5) expression. After 48 hours of doxycycline, cells were collected and total RNA was extracted for RNA sequencing. Genes differentially expressed between vehicle- and doxycycline-treated cells were identified using limma voom. For the initial analysis, filters previously demonstrated to reduce the number of false-positive calls were applied for false discovery rate (FDR <0.05) and absolute fold change (abs FC >1.5) (MAQC Consortium, 2006; SEQC/MAQC-III Consortium, 2014), producing a list of 256 up-regulated genes and 290 down-regulated genes following ELF5-V5 induction (see Additional Tables 4.1 and 4.2 at the end of this Chapter).

### Functional signatures of ELF5 overexpression

To explore the functional signatures of ELF5 overexpression, gene set enrichment analysis (GSEA) was performed using the ranked list of genes generated by limma voom as the input. GSEA statistically assesses the under- or over-representation of gene sets in the input list and can reveal the combined effects of changes in functionally related genes (Subramanian *et al.*, 2005). GSEA was performed using the C2 gene set collection from MSigDB with additional manually curated breast cancer sets (Kalyuga *et al.*, 2012). The results, shown in Figure 4.2, were visualised using the Enrichment Map plugin for Cytoscape (Merico *et al.*, 2010; Shannon *et al.*, 2003).

The Cytoscape GSEA network identifies a number of distinct hubs. Notably, the number of down-regulated gene sets in ELF5-overexpressing cells (767) far exceeds the number of up-regulated gene sets (103), using an FDR threshold of 0.05. The down-regulated hubs in the main network relate to breast cancer subtype, oestrogen response, endocrine therapy resistance, cell cycle, cancer, DNA repair, transcription, signalling, extracellular matrix and the interferon response. The up-regulated gene sets formed three main clusters relating to breast cancer subtype (including luminal, basal and claudin-low), RNA metabolism, protein translation, amino acid deprivation and signalling. Additional Figure 4.1 (located at the end of this chapter) shows enlarged images of these sub-networks with gene set names.



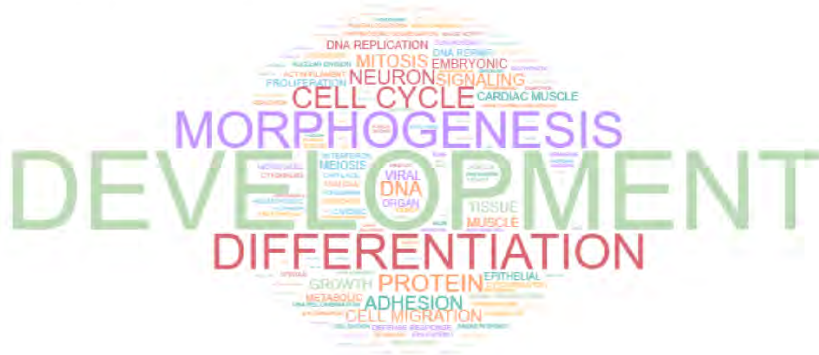
**Figure 4.2: Visualisation of the transcriptional functions of ELF5 in MCF7-ELF5-V5 cells discovered using RNA sequencing**

Cytoscape Enrichment Map visualisation of RNA-seq gene set enrichment analysis (GSEA) comparing vehicle- and doxycycline-treated MCF7-ELF5-V5 cells. GSEA was performed using the C2 gene set collection from MSigDB with additional manually curated breast

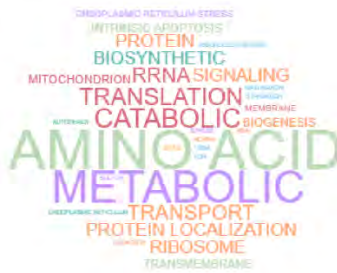
cancer sets. The circular nodes represent gene sets, with the diameter proportional to the size of the gene set. The colour of the node indicates the direction and magnitude of gene set enrichment following ELF5 induction based on normalised enrichment score (see scale), with red indicating up-regulation and blue down-regulation. Lines (“edges”) represent an overlap between connected gene sets, with the line thickness proportional to the degree of overlap. The labels summarise functional themes for prominent clusters. All gene sets with an FDR <0.05 and p-value <0.005 are shown, with the entire network containing 870 nodes and 1380 edges. Enlarged sub-networks with gene set names are shown in Additional Figure 4.1 at the end of this chapter.

GSEA was also performed using gene ontology (GO) gene sets for biological process (BP) and cellular component (CC). The BP wordclouds in Figure 4.3A and 4.3B demonstrate the occurrence of key words in the 571 down-regulated and 91 up-regulated GO BP gene sets. Once again, different functional enrichments for ELF5-driven down- and up-regulated genes were seen. The down-regulated GO BP gene sets were involved in development, differentiation, morphogenesis, cell cycle, growth and adhesion. While at first the down-regulation of development-related gene sets appears counter-intuitive, this is in fact consistent with the developmental role of ELF5, which promotes the focused differentiation of a single cell type (the alveolar secretory cell) and simultaneously suppresses alternative cell fate pathways. The smaller collection of up-regulated sets centred around metabolic, transport and translational processes, which may reflect a metabolic up-regulation in preparation for milk production. Similarly, the GO cellular component GSEA, visualised as an Enrichment Map network in Figure 4.3C, demonstrated an enrichment of down-regulated genes related to chromatin, microtubules, cytoskeleton, extracellular matrix and cell junctions, and an enrichment of up-regulated genes in ribosomal components and translational machinery. Once again, this most likely indicates an ELF5-induced change in cellular objective, with chromatin remodelling occurring to limit other developmental options, and a shift in primary function to large-scale protein (milk) production.

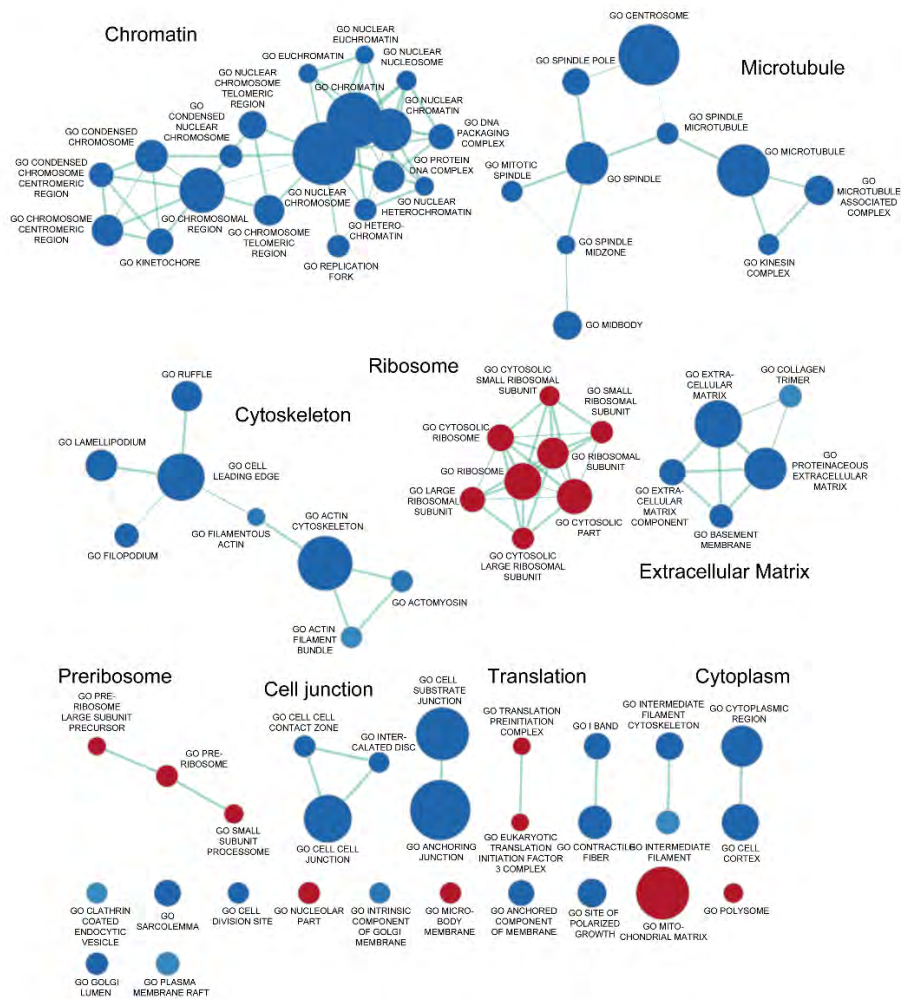
**A** GO Biological Process Down-Regulated



**B** GO Biological Process Up-Regulated



### C GO Cellular Component Gene Set Enrichment Analysis



**Figure 4.3: Gene ontology analysis of differentially expressed genes**  
(previous page)

Comparison of RNA-seq vehicle- and doxycycline-treated MCF7-ELF5-V5 cells using gene ontology (GO) biological process and cellular component gene set enrichment analysis. (A-B) Wordclouds showing key words identified in down-regulated (A) and up-regulated (B) GO biological processes. The word size is proportional to the number of occurrences (minimum 2). (C) Cytoscape Enrichment Map visualisation of enriched GO cellular component gene sets. Each circle represents a gene set, with the diameter proportional to the gene set size. The colour of the node indicates the direction and magnitude of gene set enrichment following ELF5 induction, with red indicating up-regulation and blue down-regulation. Lines (“edges”) represent an overlap between connected gene sets, with the line thickness proportional to the degree of overlap. The labels summarise functional themes for prominent clusters. All cellular component gene sets with an FDR <0.05 and p-value <0.005 are shown, with the network containing 78 nodes and 111 edges.

### **Comparison with MCF7-ELF5-V5 microarray**

A microarray study comparing doxycycline- and vehicle-treated MCF7-ELF5-V5 cells has been previously published by this laboratory (Kalyuga *et al.*, 2012). RNA sequencing has several advantages over microarrays, including the ability to detect novel transcripts, splice variants, fusion genes and single nucleotide polymorphisms (SEQC/MAQC-III Consortium, 2014). RNA-seq also has a much greater dynamic range than microarrays (which have high background and can be saturated at high expression) and can detect changes in expression of low-abundance genes more accurately (Wang *et al.*, 2014; Wang *et al.*, 2009a). Finally, RNA-seq datasets can be re-analysed as new sequences are annotated (for example, as knowledge of alternative transcripts or non-coding RNAs increases), while microarray probesets are based on current genome annotations and are fixed at the start of an experiment (Mantione *et al.*, 2014). For these reasons, RNA sequencing was chosen to provide new perspectives on the role of ELF5 in breast cancer. This represents the current gold-standard for high-sensitivity transcriptomic studies, particularly when comparing the same cell type exposed to different treatment conditions (Wang *et al.*, 2014).

Initially, the RNA-seq and microarray experiments were compared using lists of differentially expressed gene (DEG). At a false discovery rate (FDR) of 0.05, the microarray experiment produced almost twice as many DEGs as the RNA-seq experiment (Table 4.1). The introduction of a fold change filter (abs FC >1.5), aimed at reducing the number of false-positives, substantially decreased the number of DEGs in both experiments, to 547 genes in the RNA-seq experiment (a 69% decrease) and 459

in the microarray experiment (an 86% decrease). There was an obvious discrepancy between the two experiments in the proportion of up- and down-regulated DEGs, with the microarray experiment calling a greater total number and proportion of down-regulated DEGs (69.9%) compared to the RNA-seq experiment where the proportions were similar.

**Table 4.1: Numbers of Differentially Expressed Genes in MCF7-ELF5-V5 RNA-seq and Microarray Experiments**

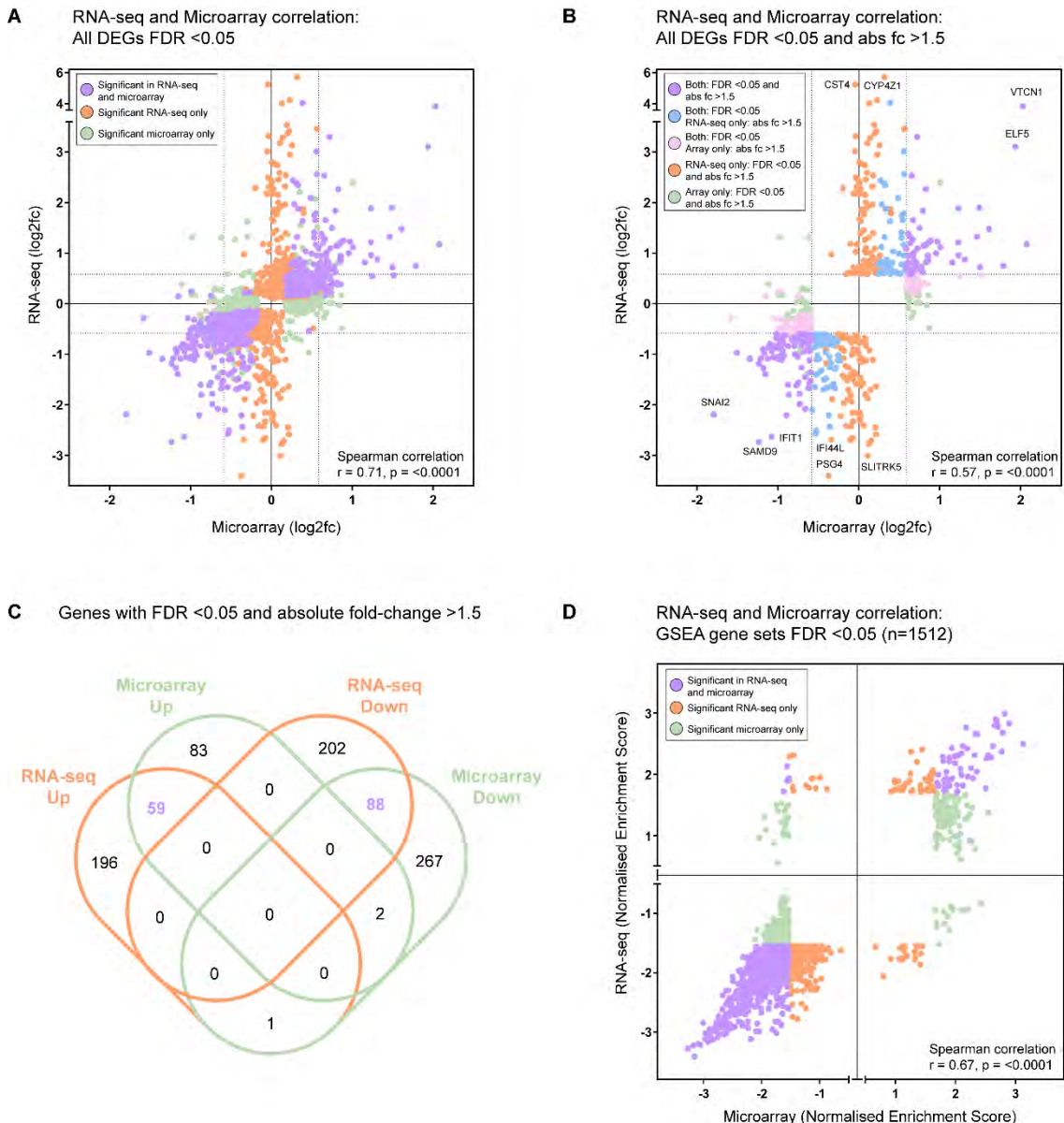
	FDR <0.05	FDR <0.05 Abs FC 1.5	FDR <0.05 Abs FC 2.0
RNA-seq up	800 (45.8%)	256 (46.9%)	145 (49.0%)
RNA-seq down	948 (54.2%)	290 (53.1%)	151 (51.0%)
RNA-seq total	1748	546	296
Microarray up	1255 (39.5%)	138 (30.1%)	21 (29.6%)
Microarray down	1921 (60.5%)	321 (69.9%)	50 (70.4%)
Microarray total	3176	459	71

Numbers of up- and down-regulated genes identified using various false discovery rate (FDR) and absolute fold change (abs FC) thresholds in the MCF7-ELF5-V5 RNA-seq (rows 1-3) and microarray (rows 4-6) experiments. Percentages (in parentheses) represent the up- or down-regulated genes numbers as a proportion of the total genes identified at the specified threshold.

Next, the correlation between the RNA-seq and microarray experiments was examined. Using all DEGs with an FDR below 0.05 (regardless of fold change), there was a strong positive correlation between the fold change values (Spearman  $r = 0.71$ ) (Figure 4.4A). There were a large number of common DEGs (shown in purple,  $n=862$ ), however there were also many genes that were unique to one experiment only (RNA-seq in orange,  $n=657$ , and microarray in green,  $n=1622$ ). It should be noted that not all genes listed in Table 4.1 were included in this analysis, which was limited to genes that were measured in both experiments (for example, the microarray does not have probes for all genes) and which had common gene names. The genes that were significant in the microarray only (green) tended to cluster around the zero mark, with only a small proportion exceeding an absolute fold change of 1.5, as indicated by the dotted line. In contrast, the genes that were significant in the RNA-seq experiment only exhibited a much wider range of changes in expression, reflecting the greater sensitivity of this method. Figure 4.4B shows only those genes with an FDR below 0.05 as well as an absolute fold change greater than 1.5 in one or both experiments (Spearman  $r = 0.57$ ).



Using the list of DEGs with an FDR <0.05 and absolute fold change >1.5, there was an overlap of 59 up-regulated genes (approximately 42% of the microarray and 23% of the RNA-seq up-regulated genes) and 88 down-regulated genes (25% of the microarray and 30% of the RNA-seq down-regulated genes) (Figure 4.4C). The 59 commonly up-regulated genes and 88 commonly down-regulated genes, represented by the purple dots in Figure 4.4B, were considered to be robustly regulated by ELF5. These genes are indicated in bold red or blue typeface in Additional Tables 4.1 and 4.2 and cluster towards the top of the FDR-sorted lists.



**Figure 4.4: Comparison of the MCF7-ELF5-V5 RNA-seq and microarray experiments**

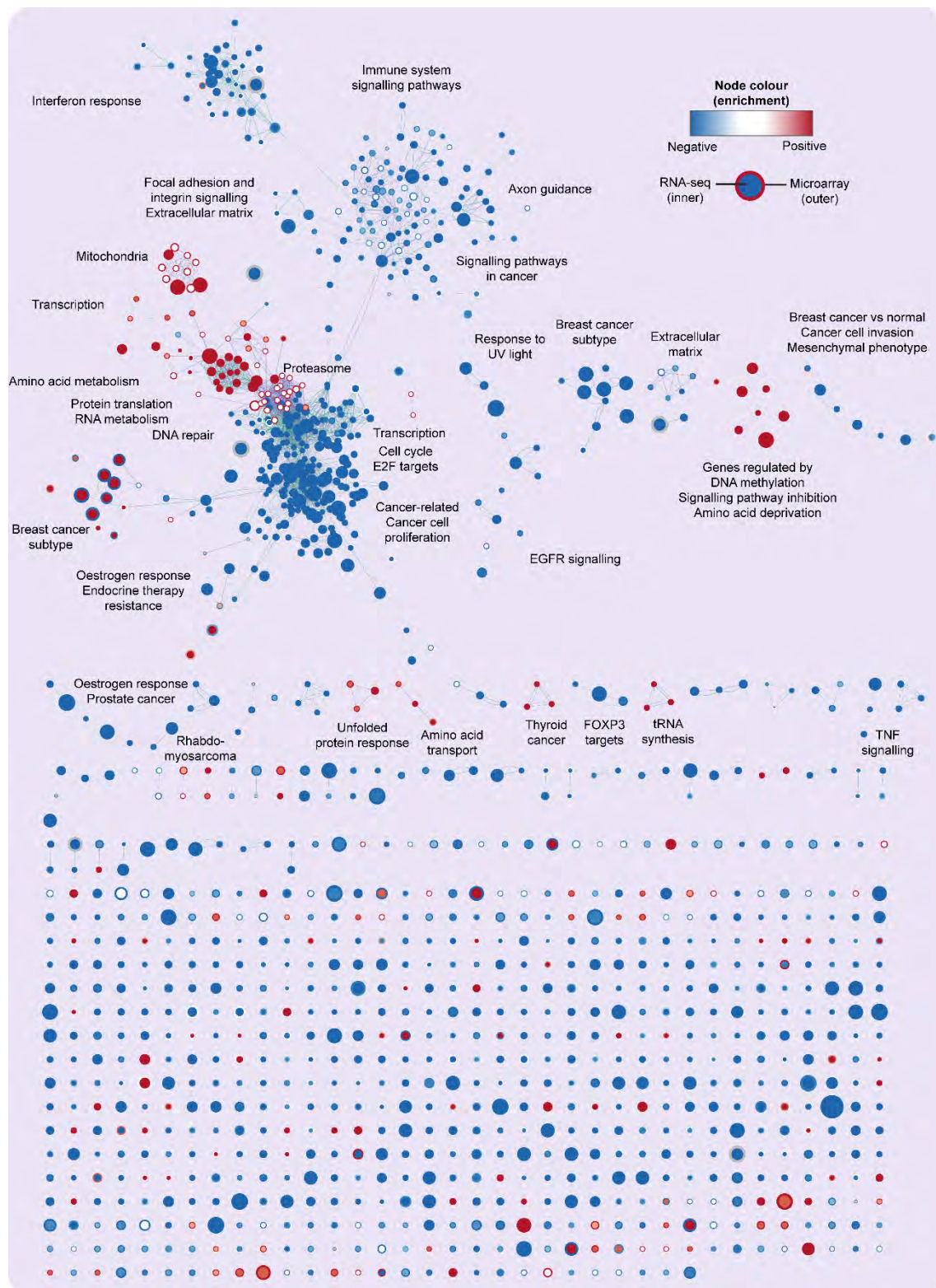
(A) Plot comparing the log2 fold change of genes with an FDR <0.05 in one or both experiments (n=3141). The microarray experiment is plotted on the x-axis and the RNA-seq experiment on the y-axis. Each dot represents a gene, with the colour indicating whether the



gene is significant at an FDR  $<0.05$  in both experiments (purple), the RNA-seq experiment only (orange) or the microarray experiment only (green). The dotted lines represent a log2 fold change level of  $\pm 0.58$ , corresponding to an absolute fold change of 1.5. The results of the Spearman rank-order correlation are shown in the bottom right corner. (B) As for panel A, with genes limited to those with an absolute fold change  $>1.5$  as well as FDR  $<0.05$  in one or both experiments. The colour of the dots indicates significance as well as fold change: purple indicates genes that meet both criteria in both experiments, orange indicates genes that meet both criteria in the RNA-seq experiment only and green indicates genes that meet both criteria in the microarray experiment only. In addition, genes that are significant at an FDR  $<0.05$  in both experiments but meet the absolute fold change criterion in only one experiment are shown in blue (absolute fold change  $>1.5$  in RNA-seq only) and pink (absolute fold change  $>1.5$  in microarray only). The names of several genes with strong fold changes are shown that were identified in both experiments (for example, ELF5, VTCN1, SNAI2), as well as genes with strong fold changes that were identified in the RNA-seq experiment only (for example, CYP4Z1, SLITRK5). (C) Venn diagram representing the overlaps of up- and down-regulated genes in the MCF7-ELF5-V5 RNA-seq and microarray experiments, using thresholds of FDR  $<0.05$  and absolute fold change  $>1.5$ . The 59 commonly up-regulated genes and 88 commonly down-regulated genes, indicated in purple, were considered to be robustly regulated by ELF5 and correspond to the purple dots in panel B. (D) Plot comparing the normalised enrichment scores for gene sets that were identified in both experiments and shown to be enriched in one or both (FDR $<0.05$ ) (n=1512). The normalised enrichment scores are derived from gene set enrichment analysis using the C2 gene set collection from MSigDB (with additional manually curated breast cancer sets). Each dot represents a gene set, with the colour indicating whether the gene is significant at an FDR  $<0.05$  in both experiments (purple), the RNA-seq experiment only (orange) or the microarray experiment only (green). The results of the Spearman rank-order correlation are shown in the bottom right corner.

Comparison with the previously published MCF7-ELF5-V5 microarray experiment was also performed by re-running GSEA on the microarray dataset using the current version of the C2 gene set collection. The correlation of the normalised enrichment scores for all gene sets with an FDR below 0.05 in one or both experiments is shown in Figure 4.4D (Spearman  $r = 0.67$ ). The majority of gene sets were enriched in the same direction in both experiments, although a small number were enriched in opposite directions (76/1512 = 5.0%). All except 4 of these oppositely enriched sets were significant in only one of the two experiments. Interestingly, 5.0% is also the FDR threshold, meaning that these discrepancies could be due to statistical error.

Next, the results of the RNA-seq and microarray GSEAs were overlapped and visualised as enrichment networks in Cytoscape (FDR  $<0.05$  and p-value  $<0.005$  in one or both experiments). The results are shown in Figure 4.5, where the inside node



**Figure 4.5: Comparison of the ELF5 transcriptional networks discovered using RNA sequencing and microarray**

Cytoscape Enrichment Map visualisation of RNA-seq gene and microarray experiments gene set enrichment analysis (GSEA) comparing vehicle- and doxycycline-treated MCF7-ELF5-V5 cells. GSEA was performed using the C2 gene set collection from MSigDB with

additional manually curated breast cancer sets. The circular nodes represent gene sets, with the diameter proportional to the size of the gene set. The node inner (RNA-seq) and outer (microarray) colour indicates the direction and magnitude of gene set enrichment based on the normalised enrichment score following ELF5 induction in each experiment, with red indicating up-regulation and blue down-regulation. Lines (“edges”) represent an overlap between connected gene sets, with the line thickness proportional to the degree of overlap; green edges represent overlaps in the RNA-seq data set, while purple edges correspond to the microarray data set. The labels summarise functional themes for prominent clusters. All gene sets with an FDR <0.05 and p-value <0.005 are shown, with the entire network containing 1219 nodes and 3306 edges.

colour represents the RNA-seq experiment (normalised enrichment score) and the node border colour represents the microarray experiment. The high level of agreement of the inner and outer node colours (more than 80% identical) once again indicates that the two networks are highly correlated. An important note is that a gene set is represented even if it is only significant (at an FDR threshold of 0.05) in one of the two experiments; the node colour reflects the normalised enrichment score for each experiment and not the significance and it is possible for a gene set to be “enriched” (according to the NES) but not statistically significant.

There were, however, some differences between the two experiments. The microarray experiment, for example, found more positively enriched gene sets relating to mitochondrial and proteasomal function. In addition, there was a small cluster of breast cancer subtype sets (19 in total) that were oppositely enriched in the two experiments; in all cases, these were up-regulated in the RNA-seq and down-regulated in the microarray. Only 3 of these sets, however, were significant (FDR <0.05) in both experiments, with relatively high FDR values (0.042-0.049) in the microarray experiment. Of the remaining 16 oppositely enriched sets, 7 were significant in the microarray experiment only and 9 were significant in the RNA-seq experiment only. However, the majority of breast cancer-related sets were identically enriched in the two experiments (82/112 at an FDR of <0.10 in a breast cancer-specific sub-analysis). Some possible reasons for this finding are discussed in further detail below.

The top three down- and up-regulated RNA-seq gene sets from the C2 GSEA are further explored in Figure 4.6. The heatmaps show the log<sub>2</sub> fold change for all genes in the set (sorted by expression in the RNA-seq experiment) and the yellow node colouring in Figure 4.6C indicates the position of these sets within the Cytoscape network. The names of all genes in the heatmaps in Figures 4.6, 4.7, and 4.8 are also

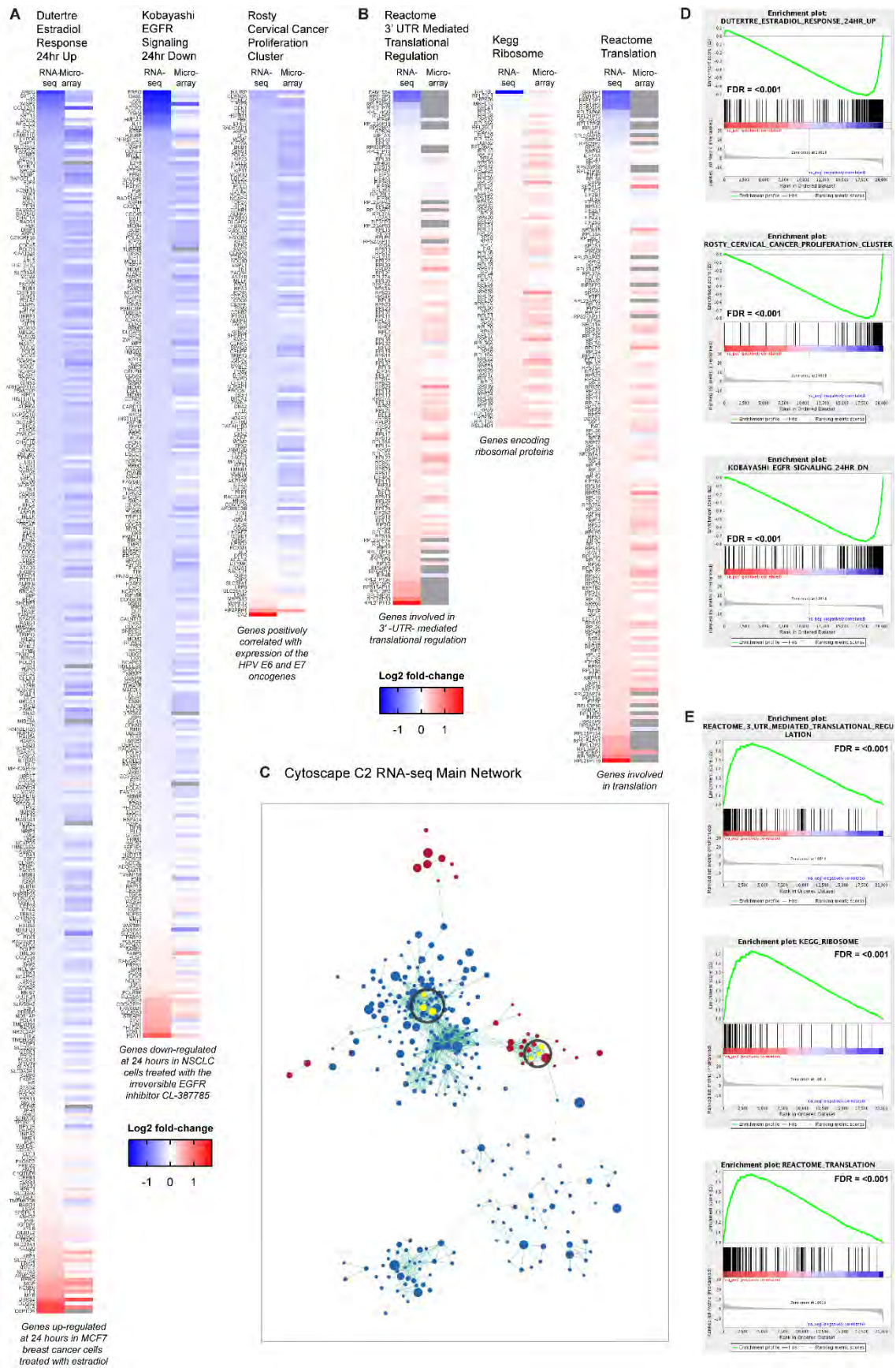
listed in Additional Tables 4.3, along with the exact log<sub>2</sub> fold-change values. The three most down-regulated gene sets (Figure 4.6A) relate to the response to oestradiol treatment, EGFR signalling and proliferation; these were also the three most down-regulated gene sets in the microarray experiment. The heatmaps demonstrate down-regulation of the majority of the genes in these sets. Conversely, strong up-regulation of gene expression can be seen in the top three up-regulated RNA-seq sets (ranked 3, 9 and 8 in the microarray experiment), which all relate to protein translation (Figure 4.6B). The corresponding enrichment plots for the top three gene sets are shown in Figure 4.6D and 4.6E; all of these sets are highly enriched, as indicated by the high maximum enrichment scores and the clear clustering of genes (represented by the vertical lines) at one end of the plot.

**Figure 4.6: Selected gene sets from the Cytoscape network**

*(next page)*

(A-B) Heatmaps representing the log<sub>2</sub> fold change (doxycycline- vs vehicle-treated MCF7-ELF5-V5 cells) for all genes in the top 3 down-regulated (A) and up-regulated (B) gene sets identified in the RNA-seq GSEA. RNA-seq data is on the left and microarray data is on the right. Rows (genes) are sorted by the magnitude of expression change in the RNA-seq experiment. Genes identified in the RNA-seq experiment that could not be automatically mapped to genes in the microarray experiment are represented by a grey box. The MSigDB gene set description is shown at the bottom of each heatmap. Gene names and fold-change values are also shown in Additional Tables 4.3A-F. (C) Representation of the RNA-seq Cytoscape GSEA network, indicating the position of the top 3 down- and up-regulated gene sets (circled yellow nodes). (D) RNA-seq GSEA enrichment plots for each of the gene sets shown in panels A and B, tracking the running enrichment score. The black bars beneath the plot indicate the position of the genes in the pre-defined set within the ranked RNA-sequencing data input list.





## Hallmark gene set enrichment analysis

MSigDB also provides a hallmark collection of gene sets, representing well-defined pathways. These sets were derived through computational overlap of related gene sets, with the aim of reducing gene set redundancy (Liberzon *et al.*, 2015). Hallmark GSEA was performed on the RNA-seq and microarray experiments and the network visualised in Cytoscape (FDR <0.10 and p-value <0.005 in one or both experiments) (Figure 4.7A). Of the 30 enriched gene sets in the network, 21 sets were significantly enriched in the same direction (up or down) in both experiments, while 8 sets were significantly enriched in one experiment only (including glycolysis, depicted as up-regulated in both by NES). Many of the commonly enriched sets validated previously described roles of ELF5 in processes such as epithelial-mesenchymal transition and cell cycle. An up-regulated metabolic signature was also identified, with up-regulation of glycolysis, adipogenesis and oxidative phosphorylation, consistent with the larger C2 Cytoscape network. Interferon alpha and gamma signalling were significantly down-regulated, with a stronger effect evident in the RNA-seq experiment compared to the microarray.

Interferon signalling was also identified as an interesting functional signature, with a significant down-regulation of both interferon alpha and interferon gamma hallmark gene sets. The genes in the interferon alpha response hallmark set are shown in the heatmap in Figure 4.7B, with the accompanying enrichment plots in Figure 4.7F (row 1). The interferon signature was also present in the microarray experiment, although the level of enrichment was not as strong; in fact, FDR values for both the alpha and gamma sets in the RNA-seq experiment were <0.001, while in the microarray experiment they were 0.075 and 0.048 respectively. In addition, interferon signalling was identified as functionally enriched by pathway analysis in the set of 291 down-regulated RNA-seq genes (FDR <0.05 and absolute fold change >1.5). In this list of 291 genes, there were 17 enriched reactome pathways, 6 of which were interferon-related and 2 of which were related to cytokine and immune signalling.

Interestingly, one gene set (oestrogen response early) was oppositely regulated in the RNA-seq (up) and microarray (down) experiments, reminiscent of the oppositely regulated breast cancer cluster in the Cytoscape network in Figure 4.5. The oestrogen response late gene set, however, was down-regulated in both conditions. Investigation of these two oestrogen response sets (containing 200 genes each) revealed a significant overlap of 101 genes. In addition, the founder sets (on which the hallmark sets are based) included all the breast cancer-related sets from the C2 gene set

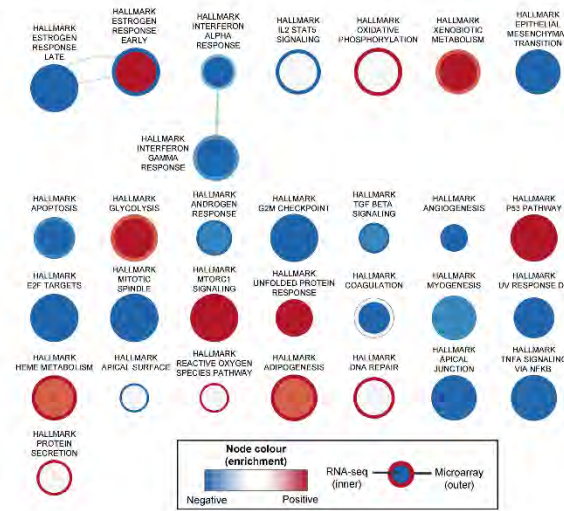
collection that were oppositely regulated in the Figure 4.5 Cytoscape network. This suggested that a small number of genes, involving the 99 non-overlapping members in each set, were likely to be responsible for the opposite regulation seen in these two hallmark sets, as well as the breast cancer cluster in the Cytoscape network.

The heatmaps in Figure 4.7C show the expression of all genes in the oestrogen response early and late gene sets in the RNA-seq and microarray experiments. In addition, microarray data from an independent study that treated hormone-deprived MCF7 cells with oestradiol for 6 hours was included (Hurtado *et al.*, 2011), demonstrating that the majority of these genes are indeed up-regulated by oestrogen in MCF7 cells. The heatmaps in Figure 4.7D show only the genes that are unique to the oestrogen response early (left) and late (right) gene sets, which overall appear very similar. It is evident from these heatmaps that the effect of ELF5 overexpression on oestrogen-regulated genes has two components - a subset of genes (for example, *KLK11*, *ACOX2*, *TRIM29*, *AREG*) are strongly down-regulated by ELF5, while another subset (for example, *AQP3*, *MSMB*, *DEPTOR*, *MUC1*) appear to be strongly up-regulated. This is also seen in one of the founder gene sets (Charafe Breast Cancer Luminal vs Basal Up) that was oppositely enriched in RNA-seq and microarray Cytoscape C2 GSEA network (Figure 4.7E). In all cases, the transition point (where genes in the set move from a negative to a positive fold change) is relatively central. This contrasts with the strongly down- and up-regulated gene set heatmaps shown in Figure 4.6.

The two components of the ELF5 gene expression response are also evident in the gene enrichment plots for the Hallmark oestrogen response, which feature clusters of genes at both ends of the expression spectrum (Figure 4.7F). In addition, the running enrichment score crosses from positive to negative near the middle of the plot, producing a curve with two (in some cases nearly equal-sized) peaks. A possible explanation for this two-component response is that it reflects the heterogeneity of single cells in the culture as they respond to cell fate cues, a hypothesis that will be explored further in the chapter discussion



## A Hallmark sets GSEA



## B Hallmark Interferon Alpha Response



## C Hallmark Estrogen Response Early



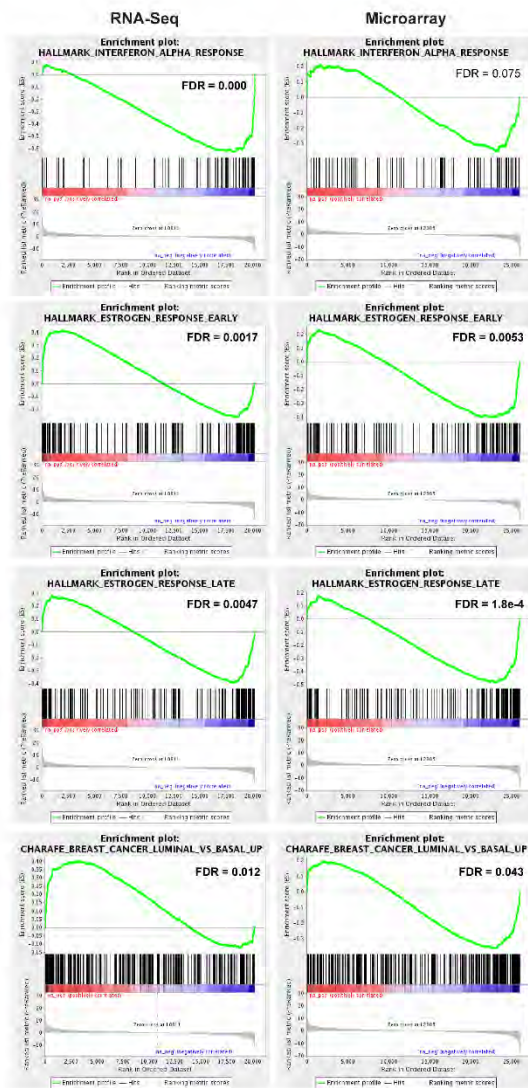
## D Estrogen Response Unique to Early



## E Charafe Breast Cancer Luminal vs Basal Up



## F





## Figure 4.7: Hallmark collection gene set enrichment analysis

(previous page)

(A) Cytoscape Enrichment Map visualisation of RNA-seq gene and microarray experiments gene set enrichment analysis (GSEA) comparing vehicle- and doxycycline-treated MCF7-ELF5-V5 cells. GSEA was performed using the Hallmark collection of gene sets from MSigDB. The circular nodes represent gene sets, with the diameter proportional to the size of the gene set. The node inner (RNA-seq) and outer (microarray) colour indicates the direction and magnitude of gene set enrichment based on normalised enrichment score following ELF5 induction in each experiment, with red indicating up-regulation and blue down-regulation. Lines (“edges”) represent an overlap between connected gene sets, with the line thickness proportional to the degree of overlap. All gene sets with an FDR <0.10 and p-value <0.005 are shown, with the network containing 30 nodes and 3 edges.

(B) Heatmap representing the log<sub>2</sub> fold change for all genes in the Hallmark Interferon Alpha Response gene set, with RNA-seq data shown on the left and microarray data on the right. Rows (genes) are sorted by the magnitude of expression change in the RNA-seq experiment. See also Additional Table 4.3G for gene names and fold-change values.

(C) Heatmaps representing the log<sub>2</sub> fold change for all genes in the Hallmark Estrogen Response Early (left) and Late (right) gene sets. Column 1 is microarray data from an independent experiment following oestradiol treatment of hormone-deprived MCF7 cells (E2 stim). Log<sub>2</sub> fold change from the MCF7-ELF5-V5 RNA-seq and microarray experiments are shown in columns 2 and 3. Rows (genes) are sorted by the magnitude of expression change in the RNA-seq experiment. Genes identified in the RNA-seq experiment that could not be automatically mapped to genes in the microarray or E2 stimulation experiments are represented by a grey box. See also Additional Tables 4.3H (early) and 4.3I (late).

(D) Heatmaps representing the log<sub>2</sub> fold change for genes unique to either the Hallmark Estrogen Response Early (left) and Late (right) gene sets (i.e. the non-overlapping genes from these two sets). See also Additional Tables 4.3J (early) and 4.3K (late).

(E) Heatmap representing the log<sub>2</sub> fold change for all genes in the Charafe Breast Cancer Luminal vs Basal Up gene set from the C2 gene set GSEA, identified as being up-regulated in the RNA-seq and down-regulated in the microarray. See also Additional Table 4.3L.

(F) GSEA enrichment plots from the RNA-seq (left) and microarray (right) experiments, tracking the running enrichment score for each of the gene sets shown in panels B, C and E.

## Oncogenic signatures gene set enrichment analysis

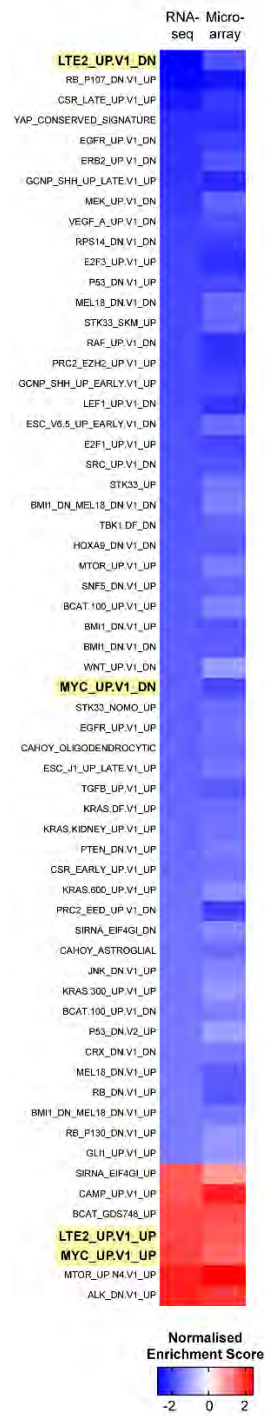
A final gene set enrichment analysis was performed using the C6 gene set collection (oncogenic signatures) from MSigDB. This collection consists of paired up- and down-regulated gene set derived from experiments targeting pathways that are commonly dysregulated in cancer. An example is the over-expression of erb-b2 receptor tyrosine kinase 2 in MCF7 cells (designed “ERBB2 UP”), which generates the up-regulated (“ERBB2\_UP.V1\_UP”) and down-regulated (“ERBB2\_UP.V1\_DOWN”) gene sets.

The significantly enriched oncogenic signature gene sets in the RNA-seq and microarray experiment are shown in Figure 4.8A. Once again, a strong correlation between the normalised enrichment scores is demonstrated (Spearman  $r = 0.79$  for the subset shown). Two oncogenic signatures were identified that: (1) Were significantly enriched in the RNA-seq experiment at an FDR  $<0.05$ , (2) Had both down- and up-regulated gene sets enriched in corresponding (i.e. opposite) directions and (3) Were also significantly enriched in the microarray dataset at an FDR  $<0.10$ .

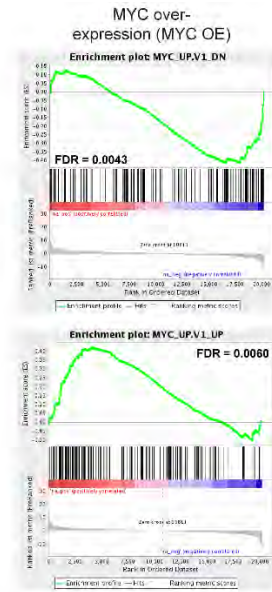
The first signature was long-term oestrogen (LTE) deprivation in MCF7 cells adapted for oestrogen-independent growth. In the down-regulated gene sets (total 55), LTE\_UP.V1\_DN was the most enriched gene set by both FDR and normalised enrichment score (NES), while in the up-regulated gene sets (total 7), LTE\_UP.V1\_UP ranked fourth. The enrichment plots for the LTE signature are shown in Figure 4.8B (column 1), with the top row demonstrating negative enrichment of LTE down-regulated genes (LTE\_UP.V1\_DN) and the bottom demonstrating positive enrichment of LTE up-regulated genes (LTE\_UP.V1\_UP); corresponding log2 fold change heatmaps for the gene sets are shown in Figure 4.8C. This is consistent with ELF5 promoting an oestrogen-insensitive phenotype, allowing cell growth and survival in the absence of oestrogenic stimulation. The genes in this list may represent candidates by which the ELF5 transcriptional network promotes endocrine resistance in luminal breast cancer.

The second signature identified was genes altered by overexpression of MYC proto-oncogene, bHLH transcription factor (MYC) in cultures of primary breast epithelial cells. There are a number of MYC-related gene sets in the C2 gene set Cytoscape network; however, these sets do not cluster together, making identification of this signature more difficult. The enrichment plots in Figure 4.8B (column 2) indicate that ELF5 overexpression promotes the expression of up-regulated MYC target genes and the repression of down-regulated MYC target genes, with the corresponding log2 fold change heatmaps shown in Figure 4.8D. A gene ontology analysis of the genes in these signature sets revealed no enrichment for genes involved in cell cycle, suggesting that these signature sets are not heavily based on MYC target genes that promote cellular proliferation (as ELF5, in contrast to MYC, has been demonstrated to acutely down-regulate proliferation). The gene set “Acosta proliferation independent MYC targets up”, for example, is up-regulated in the RNA-seq C2 gene set Cytoscape network, consistent with this hypothesis. ELF5 overexpression also resulted in increased MYC expression in the RNA-seq experiment (1.43 fold change, FDR=0.0002), although this was not observed in the microarray (1.05 fold change).

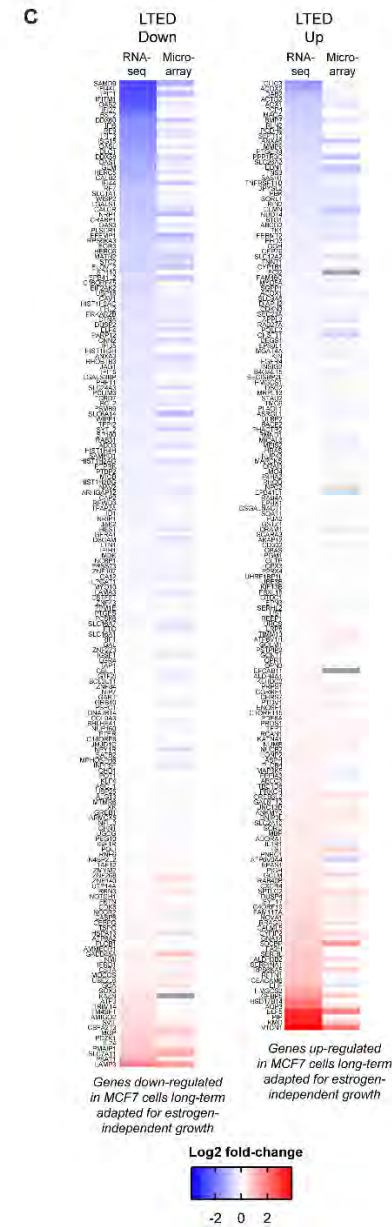
## A Oncogenic Signatures GSEA



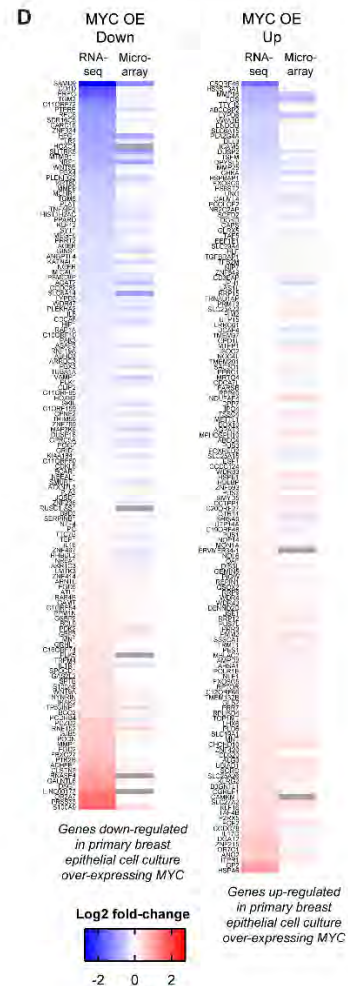
## B



## C



## D



#### **Figure 4.8: Oncogenic signatures gene set enrichment analysis**

*(previous page)*

(A) Heatmap representing the normalised enrichment score (NES) for all gene sets with an FDR <0.05 in the RNA-seq experiment, derived from gene set enrichment analysis using the C6 oncogenic signatures collection from MSigDB. RNA-seq data is shown on the left and microarray data on the right. Rows (gene sets) are sorted by the RNA-seq NES. Paired up and down gene sets that were significantly and oppositely enriched in the RNA-seq experiment as well as the microarray experiment (FDR <0.10) are highlighted in yellow.

(B) GSEA enrichment plots from the RNA-seq experiment, tracking the running enrichment score for the long-term oestrogen deprivation (LTED) gene sets (left) and MYC over-expression gene sets (right). The top row shows the enrichment for genes down-regulated by LTED or MYC, while the bottom row shows the enrichment for genes up-regulated by LTED or MYC.

(C) Heatmaps representing the log<sub>2</sub> fold change for all genes in the LTED gene sets. Rows (genes) are sorted by the magnitude of expression change in the RNA-seq experiment. Genes identified in the RNA-seq experiment that could not be automatically mapped to genes in the microarray experiment are represented by a grey box. See also Additional Tables 4.3M (down) and 4.3N (up) for gene names and fold-change values.

(D) Heatmaps representing the log<sub>2</sub> fold change for all genes in the MYC-regulated gene sets. See also Additional Tables 4.3O (down) and 4.3P (up).

#### **Enrichr analysis of ELF5-regulated genes**

As a final step in the ELF5 RNA-seq analysis, several analyses were performed using the online tool Enrichr (Chen *et al.*, 2013a; Kuleshov *et al.*, 2016). Firstly, ChIP enrichment analysis (ChEA) was performed to explore the transcription factor binding sites enriched in the promoters of input genes. In addition, enrichment of transcription factor binding motifs were analysed by Enrichr using the positional weight matrices (PWMs) from the TRANSFAC and JASPAR databases. These analyses were aimed at identifying possible ELF5-interacting transcription factors or co-factors.

Initially, the top 100 DEGs (filtered on FDR <0.05 and absolute fold change >1.5) were analysed, regardless of the direction of change. The top 10 ChIP sets (of 37 with an FDR <0.05) are shown in Figure 4.9A, revealing enrichment of ER-alpha (ESR1) and ER-beta (ESR2) and the forkhead transcription factors FOXA1 and FOXM1.

Next, the enriched ChIP binding sites and motifs were examined individually for the down-regulated (291) and up-regulated (256) RNA-seq gene sets. In the down-regulated genes, 23 ChIP sets were identified with an FDR <0.05, including known repressive proteins such as SUZ12 and zinc finger protein 217 (ZNF217) (Figure 4.9B). FOXM1 was the most enriched ChIP set in the down-regulated genes, while ChIP sets

related to hormone signalling were ranked further down and included ER-beta, GATA3 (ranked 22nd by FDR) and AR (ranked 23rd). Nine ChIP sets were enriched for binding in the promoters of the up-regulated genes, dominated by ER-alpha (ESR1 and ERA sets) and ER-beta (ESR2) (Figure 4.9C). Binding sites for GATA3, SRY-box 2 (SOX2), transcription factor E2-alpha (E2A), tumour protein 63 (p63) and cyclic AMP-dependent transcription factor ATF3 were also identified. No FOXA1 or ETS factor ChIP sets were identified in either the up- or down-regulated gene sets.

The enriched transcription factor motifs in the up-regulated and down-regulated genes were also examined. Five enriched motifs were identified in the promoters of the up-regulated genes, including three ETS family motifs (ELF3, SPI1, ETV4) (Figure 4.9D). However, in the down-regulated genes, no enriched motifs were discovered, despite the multiple ChIP-seq binding sets found to be enriched in the previous analysis. This may be explained by the use of consensus binding sequences, which do not necessarily reflect the full extent of binding in a physiological context.

One limitation of this tool is that different methods are used by the authors of the collated studies to allocate ChIP binding sites to target genes. These can be highly variable, although in many cases are likely to be based on relatively gene-proximal distance-based thresholds. Therefore, the above analyses may not be applicable if ELF5 target genes are regulated by binding to distal regulatory regions such as enhancers. The allocation of ChIP peaks to their regulated genes represents an ongoing challenge in the interpretation of ChIP-seq data (Sikora-Wohlfeld *et al.*, 2013).

Finally, the up-regulated and down-regulated RNA-seq genes were analysed for correlation with expression changes induced by perturbations in MCF7 cells (for example, treatment with a ligand or gene knockdown or over-expression) (Figure 4.9E). In the “MCF7 Up” group (that is, gene sets up-regulated by the perturbation), the most enriched set by a large margin in the RNA-seq up-regulated genes was the previously published microarray study of ELF5 overexpression (Kalyuga *et al.*, 2012). This was also the most enriched set in the “MCF7 Down” group for the RNA-seq down-regulated genes, confirming again the overall similarity between these two experiments and validating the utility of the Enrichr tool. In all groups, there were a large number of oestrogen and anti-oestrogen-related sets identified, reflecting significant positive and negative effects of ELF5 on oestrogen-regulated genes. Significant enrichment of sets relating to amino acid deprivation were also identified in the “MCF7 Up” collection for the up-regulated RNA-seq genes, consistent with the up-regulated ELF5 metabolic signature seen in the previous analyses.



**A Top 100 RNA-seq DEGs ChIP enrichment analysis (ChEA):**  
Enrichments for ChIP-seq peaks in gene regulatory region

Combined Score	Max 27.28
ESR2_21235772_ChIP-Seq_MCF-7_Human Score 27.28 FDR = 5.1e-5	
ESR1_21235772_ChIP-Seq_MCF-7_Human Score 20.50 FDR = 0.0014	
SOX2_20726797_ChIP-Seq_SW620_Human Score 20.15 FDR = 2.4e-5	
ZNF217_24962896_ChIP-Seq_MCF-7_Human Score 18.13 FDR = 5.1e-5	
ESR1_20079471_ChIP-ChIP_T-47D_Human Score 17.83 FDR = 0.0047	
AHR_22903824_ChIP-Seq_MCF-7_Human Score 16.66 FDR = 0.0010	
FOXK1_26458572_ChIP-Seq_MCF-7_Human Score 15.22 FDR = 5.1e-5	
FOXA1_27197147_ChIP-Seq_ENDOMETRIOID-ADENOCARCINOMA_Human Score 14.76 FDR = 0.0023	
TP63_17297297_ChIP-ChIP_HaCaT_Human Score 14.38 FDR = 0.024	
ARNT_22903824_ChIP-Seq_MCF-7_Human Score 12.40 FDR = 0.0025	

**B RNA-seq Down-regulated DEGs ChIP enrichment analysis (ChEA):**  
Enrichments for ChIP-seq peaks in gene regulatory region

Combined Score	Max 38.71
FOXK1_26458572_ChIP-Seq_MCF-7_Human Score 38.71 FDR = 1.5e-11	
SUZ12_18555785_ChIP-Seq_MESCs_Mouse Score 23.70 FDR = 1.7e-5	
ESR2_21235772_ChIP-Seq_MCF-7_Human Score 22.04 FDR = 2.9e-4	
ZNF217_24962896_ChIP-Seq_MCF-7_Human Score 19.69 FDR = 2.2e-5	
SUZ12_18592474_ChIP-Seq_MEFs_Mouse Score 18.37 FDR = 9.9e-5	
SUZ12_18592474_ChIP-Seq_MESCs_Mouse Score 17.65 FDR = 6.1e-5	
SUZ12_18974828_ChIP-Seq_MESCs_Mouse Score 16.42 FDR = 7.6e-5	
MTF2_20144788_ChIP-Seq_MESCs_Mouse Score 15.58 FDR = 5.8e-6	
CJUN_26782858_ChIP-Seq_BT549_Human Score 14.80 FDR = 7.7e-5	
SOX2_20726797_ChIP-Seq_SW620_Human Score 14.03 FDR = 4.7e-4	

**C RNA-seq Up-regulated DEGs ChIP enrichment analysis (ChEA):**  
Enrichments for ChIP-seq peaks in gene regulatory region

Combined Score	Max 38.56
ESR1_21235772_ChIP-Seq_MCF-7_Human Score 38.56 FDR = 5.6e-6	
ESR2_21235772_ChIP-Seq_MCF-7_Human Score 28.60 FDR = 3.2e-5	
ESR1_20079471_ChIP-ChIP_T-47D_Human Score 27.96 FDR = 2.7e-4	
SOX2_20726797_ChIP-Seq_SW620_Human Score 11.9 FDR = 0.0017	
ERA_27197147_ChIP-Seq_ENDOMETRIOID-ADENOCARCINOMA_Human Score 7.04 FDR = 0.045	
GATA3_24758297_ChIP-Seq_MCF-7_Human Score 5.57 FDR = 0.020	
P63_26484246_ChIP-Seq_KERATINOCYTES_Human Score = 5.14 FDR = 0.035	
E2A_27217539_ChIP-Seq_RAIOGS-Cell_line_Human Score = 4.79 FDR = 0.045	
ATF3_27145783_ChIP-Seq_COLON_Human Score = 4.65 FDR = 0.045	

**D RNA-seq Up-regulated DEGs:**  
Enrichments for transcription factor motifs in regulatory region

Combined Score	Max 7.63
ELF3 (human) Score 7.63 FDR = 0.012	
NR5A2 (human) Score 7.27 FDR = 0.010	
CEBPB (human) Score 6.45 FDR = 0.038	
SP1 (human) Score 5.88 FDR = 0.038	
ETV4 (human) Score 5.50 FDR = 0.038	

**E Enrichments for MCF7 Perturbations**

Enrichments for MCF7 Perturbations

ELF5 effects mimicking perturbation or ligand

ELF5 effects opposing perturbation or ligand

Up-regulated genes

Combined Score

Max 250.0

ELF5 overexpression\_GSE30405\_mcf7:423 Score 250.0 FDR = 3.3e-61

Trans-retinoic acid\_GSE32161\_mcf7:19 Score 50.12 FDR = 2.4e-13

Glutamine deprivation\_GSE62673\_mcf7:159 Score 45.91 FDR = 3.3e-12

Cimicifuga racemosa extract\_GSE6800\_mcf7:468 Score 40.51 FDR = 1.7e-11

Serine deprivation\_GSE62673\_mcf7:170 Score 38.67 FDR = 4.5e-11

Trans-retinoic acid + hydrocorti\_GSE32161\_mcf7:20 Score 34.15 FDR = 1.5e-9

SPDEF shRNA\_GSE40985\_mcf7:137 Score 33.49 FDR = 1.5e-9

Resveratrol\_GSE25412\_mcf7:310 Score 32.20 FDR = 4.4e-9

5GY radiation\_GSE59734\_mcf7:477 Score 31.44 FDR = 7.1e-9

Isoleucine deprivation\_GSE62673\_mcf7:161 Score 29.56 FDR = 1.2e-8

Down-regulated genes

Combined Score

Max 143.4

SPDEF shRNA\_GSE40985\_mcf7:137 Score 143.4 FDR = 2.6e-37

TSC2 shRNA\_GSE88324\_mcf7:211 Score 108.6 FDR = 7.2e-28

Knockdown of ESR1\_GSE18431\_mcf7:326 Score 84.19 FDR = 6.3e-23

3beta-Adiol(17beta-diol)\_GSE33287\_mcf7:331 Score 73.9 FDR = 3.7e-19

Cyclin D1 siRNA\_GSE48989\_mcf7:344 Score 71.13 FDR = 1.4e-19

AP-2 gamma siRNA\_2\_GSE15481\_mcf7:50 Score 68.96 FDR = 7.5e-18

3beta-Adiol(17beta-diol)\_GSE33287\_mcf7:330 Score 67.94 FDR = 1.7e-18

AP-2 gamma targeting siRNA\_GSE15481\_mcf7:470 Score 66.37 FDR = 2.7e-17

Knockdown of NR2E3\_GSE18431\_mcf7:325 Score 65.87 FDR = 2.3e-18

Ethanol and 27-Hydroxycholesterol\_GSE46924\_mcf7:9 Score 57.55 FDR = 5.8e-15

MCF7 Up

Combined Score

Max 42.35

Estradiol\_GSE33287\_mcf7:328 Score 42.35 FDR = 5.5e-12

3beta-Adiol(17beta-diol)\_GSE33287\_mcf7:331 Score 35.03 FDR = 5.3e-10

Estrogen\_GSE11324\_mcf7:59 Score 25.90 FDR = 7.4e-7

E2 100 nM for 10 hrs\_GSE24085\_mcf7:491 Score 24.52 FDR = 7.4e-7

Doxycycline\_GSE37945\_mcf7:364 Score 21.55 FDR = 1.7e-6

Estradiol\_GSE33287\_mcf7:329 Score 21.52 FDR = 2.3e-6

Estradiol\_GSE45643\_mcf7:317 Score 20.88 FDR = 5.0e-6

3beta-Adiol(17beta-diol)\_GSE33287\_mcf7:330 Score 20.45 FDR = 2.3e-6

HOXB7 overexpression\_GSE63607\_mcf7:366 Score 19.36 FDR = 4.8e-6

17beta-estradiol\_GSE20081\_mcf7:130 Score 19.17 FDR = 1.0e-5

MCF7 Down

Combined Score

Max 322.3

ELF5 overexpression\_GSE30405\_mcf7:423 Score 322.3 FDR = 2.2e-84

Trans-retinoic acid + hydrocorti\_GSE32161\_mcf7:20 Score 107.5 FDR = 1.9e-27

Trans-retinoic acid\_GSE32161\_mcf7:19 Score 98.27 FDR = 3.6e-25

Erlotinib\_GSE30516\_mcf7:126 Score 80.53 FDR = 6.8e-22

Estradiol\_GSE46924\_mcf7:467 Score 69.11 FDR = 1.8e-17

siRNA\_GSE31118\_mcf7:14 Score 66.62 FDR = 4.5e-18

100nM BI2536 for 1hr and E2 for\_GSE46856\_mcf7:231 Score 63.40 FDR = 3.8e-18

14-3-3 overexpression\_GSE37139\_mcf7:434 Score 63.37 FDR = 2.0e-18

Bortezomib (Velcade) + Estrogen\_GSE31118\_mcf7:13 Score 61.67 FDR = 7.7e-17

HOXB7 overexpression\_GSE63607\_mcf7:366 Score 59.98 FDR = 2.7e-17

**Figure 4.9: “Enrichr” analysis of RNA-seq-identified differentially expressed genes**

(A) Enriched ChIP sets (ranked by Enrichr combined score) identified in the regulatory regions of the top 100 differentially expressed MCF7-ELF5-V5 RNA-seq genes (filtered for absolute fold change >1.5 and ranked by FDR). The identifier for each ChIP set contains the name of the transcription factor followed by the PubMed ID, the type of experiment (ChIP-seq or ChIP-chip), the cell line or tissue, and the species. The top 10 sets (of 37 sets with an FDR <0.05) are shown. Analysis was performed using the Enrichr ChIP enrichment analysis (ChEA) tool. The combined score calculated by Enrichr represents a combination of the p-value (Fisher’s exact test) and z-score (reflecting deviation from an expected rank).

(B) Enriched ChIP sets (ranked by Enrichr combined score) identified in the regulatory regions of down-regulated genes in the MCF7-ELF5-V5 RNA-seq experiment, defined by FDR <0.05 and absolute fold change 1.5. The top 10 sets (of 23 sets with an FDR <0.05) are shown.

(C) Enriched ChIP sets (ranked by Enrichr combined score) identified in the

regulatory regions of up-regulated genes in the MCF7-ELF5-V5 RNA-seq experiment, defined by FDR <0.05 and absolute fold change 1.5. A total of 9 ChIP sets with an FDR <0.05 were identified (all shown). (D) Enriched transcription factor motifs from the TRANSFAC and JASAPR databases, analysed using Enrichr, identified in the regulatory regions of up-regulated genes in the MCF7-ELF5-V5 RNA-seq experiment. A total of 5 enriched motifs with an FDR <0.05 were identified (all shown). No enriched motifs with an FDR <0.05 were identified for the down-regulated RNA-seq genes. (E) Enriched gene sets from the GEO database, identified using Enrichr, related to perturbations in MCF7 cells (for example, treatment with a ligand or gene knockdown/over-expression). “MCF7 up” indicates that the perturbation up-regulates genes in the set (top row), while “MCF7 down” indicates that the perturbation down-regulates genes in the set. The top 10 MCF7 gene set overlaps with significantly up-regulated genes (left) and down-regulated genes (right) from the MCF7-ELF5-V5 RNA-seq experiment, defined as above, are shown. Enriched sets are ranked by the combined score and each set is labelled with the GEO database reference, combined score and FDR (<0.05 in all cases).

### **Identification of ELF5 genomic binding sites in MCF7-ELF5-V5 cells using ChIP-seq**

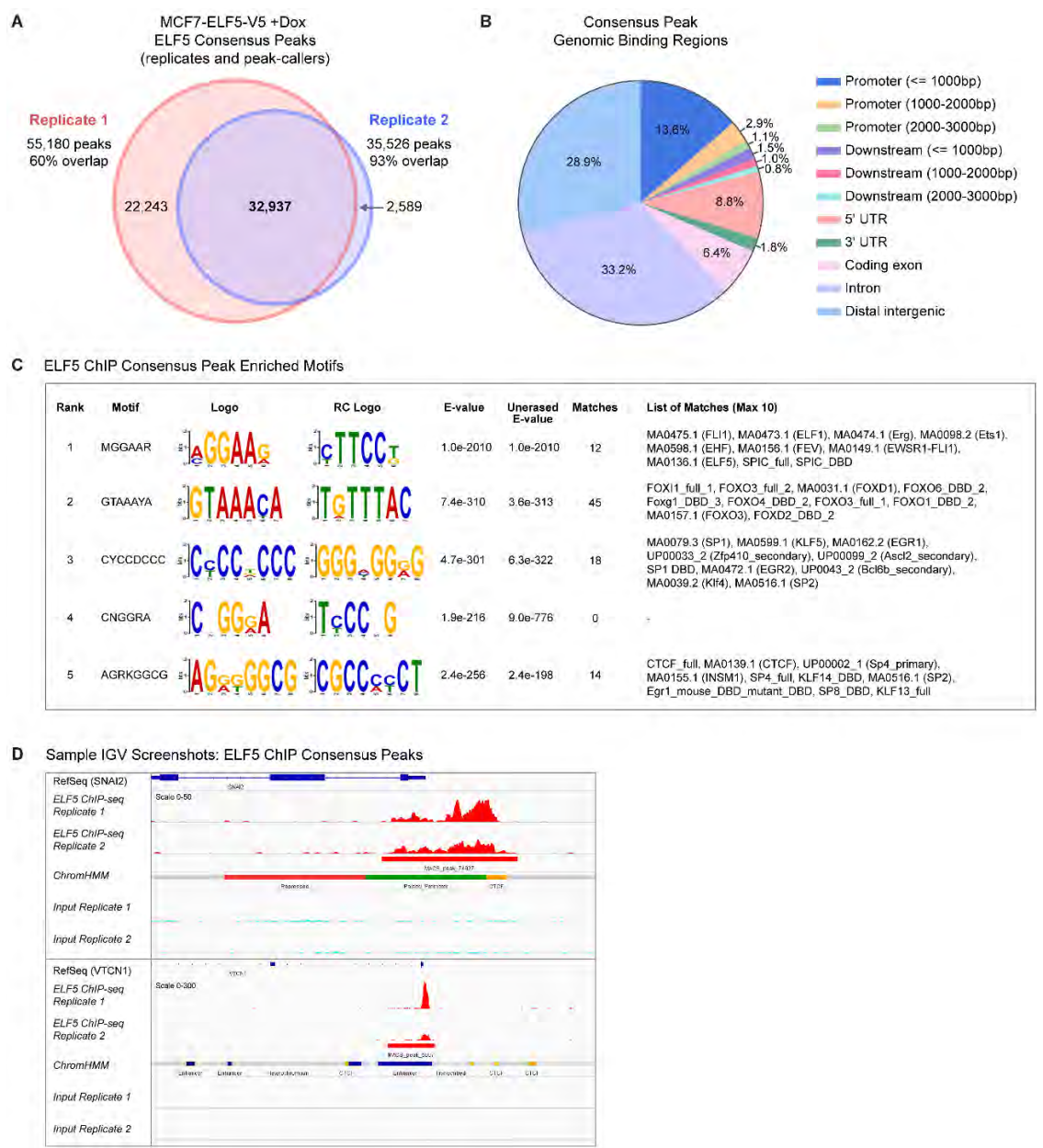
To study the ELF5 transcriptome in further detail, ELF5 ChIP-seq was performed using MCF7-ELF5-V5 cells that had been treated with doxycycline for 48 hours. A combination of ELF5 and V5 antibodies was used to extract ELF5-V5-bound DNA, which was sequenced in two independent replicates. ELF5 binding regions were identified using two peak calling programs (MACS and HOMER). There were 32,937 peaks that were called by MACS and HOMER in both replicates, forming the consensus set of ELF5 ChIP-seq peaks (Figure 4.10A).

Analysis of the genomic regions of ELF5 binding using the *Cis*-regulatory Element Annotation System (CEAS) tool revealed a significant enrichment for promoter regions (with 17.6% of all binding sites located within 3000 base pairs of the transcriptional start site of a gene), as well as the 5' untranslated region (UTR) (8.6%). Intronic and distal intergenic binding sites were decreased compared to the overall genomic distribution, although these regions still accounted for 62.1% of all ELF5 binding sites (Figure 4.10B).

*De novo* motif analysis was performed to identify enriched genomic sequences in the ELF5 consensus peaks. As expected, the most significantly enriched motif consisted of a core GGAA sequence, matching multiple ETS family database members including ELF5 (Figure 4.10C). Interestingly, the second-most enriched motif was a Forkhead motif, matching multiple members of the Forkhead family including FOXA1, hinting at a

possible connection between these two transcription factors. Other significant motifs identified in the top 5 included specificity protein (SP) transcription factors, Kruppel like factors (KLF), CCCTC-binding factor (CTCF) and early growth response protein 1 (EGR1), all of which are members of the zinc finger transcription factor family.

Two examples of ELF5 ChIP-seq consensus peaks are shown in Figure 4.10D. These peaks are located upstream from the transcriptional start sites of two genes shown to be significantly down-regulated (*SNAI2*) or up-regulated (*VTCN1*) by MCF7-ELF5-V5 RNA-seq. In addition, chromatin states for MCF7 cells based on histone modifications (from Taberlay *et al.*, 2014) are shown in the ChromHMM track, demonstrating ELF5 binding at poised promoter and enhancer regions in these examples.





#### Figure 4.10: ELF5 ChIP-seq summary

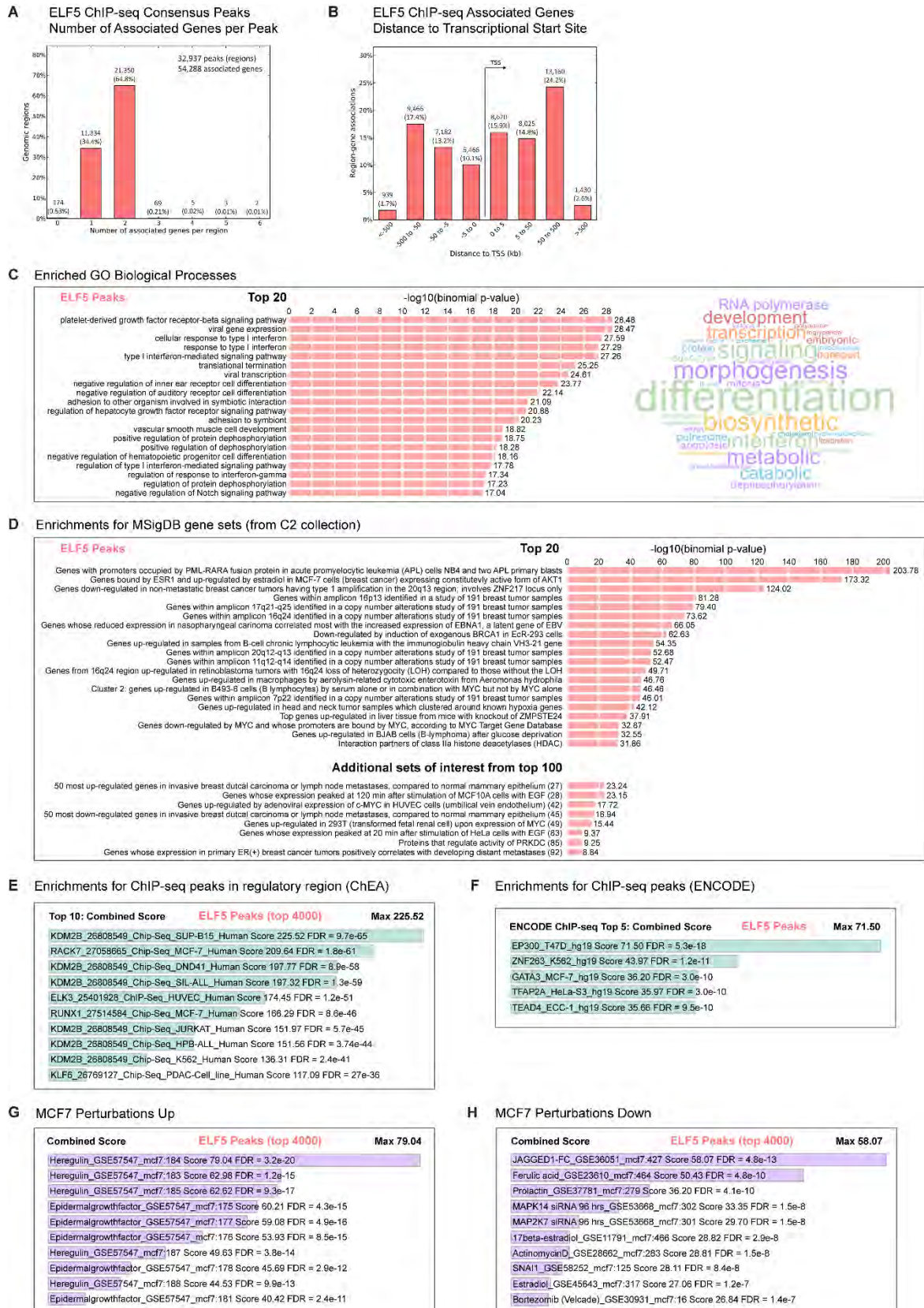
(previous page)

(A) Numbers of ELF5 binding sites identified by both peak-callers (MACS and HOMER) in each replicate, overlapped to produce a consensus set of 32,927 ELF5 binding sites. (B) Genomic binding regions of the 32,937 ELF5 consensus peaks, generated using the *Cis*-regulatory Element Annotation System (CEAS) tool. (C) *De novo* motif analysis of the 32,937 ELF5 consensus peaks using DREME and Tomtom. The top 5 motifs identified are represented as forward and reverse logos. The E-value, an indicator of statistical significance, represents the Fisher's exact test p-value multiplied by the number of candidate motifs tested. The best database matches for the identified motifs (up to a maximum of 10) are shown on the right. The most enriched motif is an ETS motif, followed by a Forkhead motif. (D) Integrative Genomics Viewer (IGV) screenshots showing two examples of ELF5 binding at consensus peaks, located upstream of the transcriptional start sites of the *SNAI2* (top) and *VTCN1* (bottom) genes. Chromatin states for MCF7 cells based on histone modifications (from Taberlay *et al.*, 2014) are shown in the ChromHMM track, demonstrating ELF5 binding at poised promoter and enhancer regions in these examples. Minimal ChIP-seq signal is seen in the input control samples (bottom two tracks).

#### GREAT functional analysis of ELF5 ChIP-seq consensus peaks

Functional analysis of the 32,937 ELF5 ChIP-seq consensus peaks was performed using the Genomic Regions Enrichment of Annotations Tool (GREAT) (McLean *et al.*, 2010). This tool attempts to address several problems associated with common methods of associating *cis*-regulatory regions with target genes; these include the loss of a significant number of binding events (when only those events proximal to transcriptional start sites are considered) or a bias towards genes that are flanked by large intergenic regions (when events are assigned to the closest one or two genes). GREAT assigns a default regulatory domain to each gene, consisting of a basal region (5kb upstream and 1kb downstream of the transcriptional start site) and an extension up to the basal regulatory domain of the nearest upstream and downstream genes within 1 Mb. Each genomic region (for example, ChIP-seq peak) is then associated with all genes with which there is a regulatory domain overlap.

Using this approach, 54,288 genes (15,603 unique) were found to be associated with the 32,937 ELF5 ChIP-seq consensus peaks, with most peaks associated with 1-2 genes (Figure 4.11A). The distribution of the peaks around the transcriptional start site (TSS) of the associated genes is shown in Figure 4.11B, demonstrating approximately 10% of genes with a peak within 5kb upstream and 16% within 5kb downstream of the gene TSS. A significant proportion of genes (almost 70%) were assigned to peaks that were distributed at distances of 5-500kb upstream or downstream of the TSS.



**Figure 4.11: Functional analysis of ELF5 ChIP-seq peaks**

Genomic Regions Enrichment of Annotations Tool (GREAT) (A-D) and Enrichr (E-H) analyses of the 32,937 ELF5 ChIP-seq consensus peaks. (A) Number of associated genes

per consensus peak, based on default GREAT settings. There are 54,288 non-unique (and 15,603 unique) genes associated with the 32,937 peaks, with the majority of peaks associated with 1-2 genes. (B) Peak-associated genes grouped by distance of the consensus peak to the transcriptional start site (TSS). (C) Top 20 enriched GO Biological Processes (BP) generated by GREAT and ranked by  $-\log_{10}(\text{binomial p-value})$ . FDR of all sets is  $<0.05$ . The associated wordcloud is based on the top 100 GO BP sets (FDR  $<0.05$ ), with word size proportional to the number of occurrences (minimum 2). (D) Enriched gene sets from MSigDB C2 gene set collection (chemical and genomic perturbations subset) generated by GREAT and ranked by  $-\log_{10}(\text{binomial p-value})$ . The top 20 sets are shown, as well as additional sets of interest from the top 100 (rank shown in parentheses after gene set title). Descriptive names for gene sets are shown (in contrast to the abbreviated names shown in previous Cytoscape RNA-seq GSEA figures). FDR of all sets is  $<0.05$ . (E) Enrichr analysis of the top 4,000 ELF5 consensus peaks, ranked by MACS score. Each peak was allocated to genes according to the Enrichr algorithm and capped at a maximum of 2,000 genes. The top 10 enriched ChIP sets identified in the regulatory regions of the associated genes are shown, sorted by Enrichr combined score. The identifier for each ChIP set contains the name of the transcription factor followed by the PubMed ID, the type of experiment (ChIP-seq or ChIP-chip), the cell line or tissue, and the species. (F) Enriched ENCODE ChIP-seq sets identified in the regulatory regions of the Enrichr peak-associated genes (top 5). (G-H) Enriched gene sets from the GEO database, identified by Enrichr analysis of the top 4,000 ELF5 consensus peaks ranked by MACS score and capped at 2000 associated genes. Gene sets are related to perturbations in MCF7 cells (for example, treatment with a ligand or gene knockdown/overexpression). “MCF7 Perturbations Up” indicates that the perturbation up-regulates genes in the set (G), while “MCF7 Perturbations Down” indicates that the perturbation down-regulates genes in the set (H). Enriched sets are ranked by the combined score and each set is labelled with the GEO database reference, combined score and FDR ( $<0.05$  in all cases).

GREAT analysis revealed enrichment of a number of gene ontology (GO) biological processes (Figure 4.11C). Several of the most enriched processes were related to the viral response and interferon signalling, with others related to growth factor signalling, translation, differentiation and Notch signalling. Several pathways were highlighted by this analysis, including interferon signalling (which was also identified as a strong down-regulated pathway in the RNA-seq GSEA), platelet-derived growth factor signalling and Notch signalling. These last two pathways were also identified in the RNA-seq GSEA but were less obviously clustered than interferon signalling; this demonstrates that analysis of ChIP-seq peaks can help to refine the results of GSEA for ongoing studies. In addition, this analysis demonstrates the intersection of cell-intrinsic (related to differentiation) and cell-extrinsic (related to communication with the

extracellular environment and immune system) functions of ELF5 (Gallego-Ortega *et al.*, 2015; Kalyuga *et al.*, 2012).

The results from the top 100 enriched biological processes are represented as a wordcloud to the right of the chart. Overall, the enriched biological processes appear broadly similar to those identified in the up- and down-regulated RNA-seq gene sets and again demonstrate the cell-intrinsic (differentiation, morphogenesis, metabolic) and cell-extrinsic (interferon, signalling) functions. This suggests that ELF5 binding directly regulates the expression of genes responsible for these processes. Despite the general similarity, the overlap of specific biological processes identified through GREAT ChIP-seq peak analysis and gene set enrichment analysis of RNA-seq data was only 5% (using a maximum of 100 significant biological processes ranked by FDR). This may be related to various factors, including the fold change threshold used in the RNA-seq data or the method used by GREAT to assign ChIP peaks to target genes. In addition, it is likely that not all ELF5 binding sites results in changes in gene expression, as has been demonstrated in other ChIP-seq studies (Shlyueva *et al.*, 2014). In addition, ELF5 peaks associated with cell-extrinsic functions may not promote changes in expression, as MCF7 cells in culture lack the interactions with the extracellular microenvironment that may be required for full activation of these pathways.

GREAT was also used to identify enriched gene sets from the MSigDB database (C2 chemical and genomic perturbations, representing a subset of those used for the RNA-seq C2 GSEA). In the top 20 sets (ranked by p-value), 8 sets were identified that were directly related to breast cancer, with one set specifically relating to ESR1-regulated genes and another to ZNF217-regulated genes (Figure 4.11D). Two additional sets within the top 20 are related to MYC target genes. Several additional sets from the top 100 (FDR <0.05) are also shown, relating to breast cancer (3 sets), growth factor stimulation (2 sets) and MYC action (2 sets). A gene set related to proteins regulating PRKDC (also known as DNA-PKcs) was also identified in the top 100, which is relevant given the newly-discovered interaction between ELF5 and DNA-PKcs discussed in Chapter 5.

### **Enrichr functional analysis of ELF5 ChIP-seq peaks**

The top 4,000 ELF5 ChIP-seq consensus peaks (ranked by MACS score) were also analysed using Enrichr, which automatically assigns each peak to the closest gene, with a maximum of 2,000 genes (Chen *et al.*, 2013a; Kuleshov *et al.*, 2016). There were a large number of additional ChIP-seq factors that were identified as being

enriched in the promoters of the 1,655 unique associated genes (Figure 4.11E, ChIP enrichment analysis or ChEA). Lysine-specific demethylase 2B (KDM2B), which is primarily responsible for demethylating H3K4me3 at gene promoters, was identified in 6 of the 10 most-enriched ChIP-seq sets, suggesting KDM2B as a possible ELF5 co-factor. Two additional repressive proteins previously identified in the Enrichr analysis of the RNA-seq data were the transcription factor ZNF217 and polycomb complex member SUZ12; ZNF217 was once again identified in the ELF5 ChIP-seq peak analysis (FDR =  $1.0 \times 10^{-16}$ , ranked 23 by combined score), while the polycomb complex member SUZ12 was not significantly enriched. Interestingly, binding of the ETS transcription factor ELK3 was highly significant (FDR =  $1.2 \times 10^{-51}$ ), while ELF5 ChIP-seq in T47D cells ranked 13 (FDR =  $5.0 \times 10^{-19}$ ). Several FOXA1 and ESR1 datasets were also identified as significantly enriched.

In addition, binding of the active enhancer-associated protein histone acetyltransferase protein p300 was enriched in the ELF5 ChIP-seq peaks (FDR =  $1.4 \times 10^{-25}$ , ranked 16 by combined score in the ChEA analysis). In the ENCODE ChIP-seq dataset, p300 was the most enriched set by a significant margin. This indicates a potentially important role for ELF5 at enhancer regions, which is likely to be broader than this analysis (which is limited to a subset of peaks at relatively proximal regulatory regions) suggests.

Enrichr was also used to examine the peak-associated genes for enrichments in the MCF7 perturbation gene sets. There was a striking enrichment for growth factor up-regulated genes in the ELF5 ChIP-seq associated genes, primarily related to epidermal growth factor receptor ligands including EGF and heregulin (Figure 4.11G). The RNA-seq data suggests that ELF5 binding, at least in the short-term, is primarily acting to inhibit the expression of growth factor-regulated genes (for example, the down-regulation of EGFR signalling in Figure 4.6A). These peaks may also be priming the cell for future activation of these genes in response to oestrogen-independent growth signals. In the MCF7 down perturbation sets, there was a more varied enrichment of gene sets. The top ten gene sets included prolactin treatment, oestrogen treatment and cytokine-related signalling pathways (MAP kinase silencing). The most enriched gene set in this category was treatment with the Notch pathway ligand Jagged-1 (JAG1), consistent with a previously described role for ELF5 in suppressing the mammary stem cell-promoting Notch pathway (Chakrabarti *et al.*, 2012b).

## Identification of the direct regulatory targets of ELF5

A probable set of direct ELF5 target genes was generated by integrating the RNA-seq gene expression and ChIP-seq data. For this analysis, each ChIP-seq peak was assigned to the closest gene to form a peak-gene list; this was then filtered to include only those genes where the ChIP-seq peak was within 10kb of the transcription start site (TSS) and then overlapped with the list of RNA-seq differentially expressed genes. This method has been previously used for analysis of ELF5 ChIP-seq data (Kalyuga *et al.*, 2012) and involves stricter criteria for allocating peaks to genes than the GREAT method described above. As almost 20% of ELF5 ChIP-seq peaks are found in promoters (<3kb), this method is likely to detect most genes that are regulated by ELF5 binding to the promoter and other regulatory regions within 10kb (i.e. high specificity). It does, however, mean that approximately 50% of all the peaks in the dataset are not included in the peak-gene analysis. In addition, this method will not detect target genes regulated by more distal enhancers. The choice of method for peak-gene allocation therefore represents a compromise between specificity and sensitivity.

Overlap of the peak-gene list with the RNA-seq DEGs identified 357 genes as probable ELF5 direct targets, representing approximately 72% of the total DEGs as defined by the previous criteria (FDR <0.05 and absolute fold change >1.5) (Figure 4.12A). Of these 357 genes, the distribution between up- and down-regulated genes was relatively even, at 176 and 181 genes respectively.

Functional analyses of the up- and down-regulated direct target genes were performed using the MSigDB Biological Process and Hallmark gene sets (Figures 4.12B and 4.12C). These analyses demonstrate that many of the canonical ELF5 functions, such as differentiation, epithelial-mesenchymal transition, oestrogen response, interferon response and metabolic up-regulation, are likely to be directly regulated by ELF5 binding at relatively proximal sites. Interestingly, despite the comparable numbers of up- and down-regulated target genes identified, the functional enrichments were much more significant in the down-regulated targets. This suggests that, as a whole, the down-regulated target genes are more functionally similar to each other than the up-regulated genes.

The hallmark sets oestrogen response early and late were again identified in both the up- and down-regulated gene sets, with oestrogen response early the most enriched set for the up-regulated genes by a significant margin (overlap 14 ELF5 genes with the 200 genes in the set). Cell fate dynamics and single cell heterogeneity, as suggested

previously, may contribute to this two-component response. The presence of ELF5 binding to the promoters of these genes, however, suggests a second possibility - that ELF5 and ER may regulate a subset of specific genes in an identical manner. When the 14 ELF5 up-regulated “oestrogen response early” genes are analysed for enriched biological processes, significant results include the regulation of molecule and ion transport (6 genes), regulation of cell differentiation (5 genes) and negative regulation of cell proliferation (4 genes). ELF5 and ER may therefore promote the expression of a subset of genes that are important for epithelial cell identity. What it is not possible to know from these experiments, however, is whether the cells with increased expression of these ER-induced genes are the same cells that have ELF5 binding in the promoters of these genes. Due to the single-cell heterogeneity that is almost inevitably present in the culture, it is possible that these events (up-regulation of ER-induced genes and ELF5 promoter binding) are occurring in different sub-populations of cells.

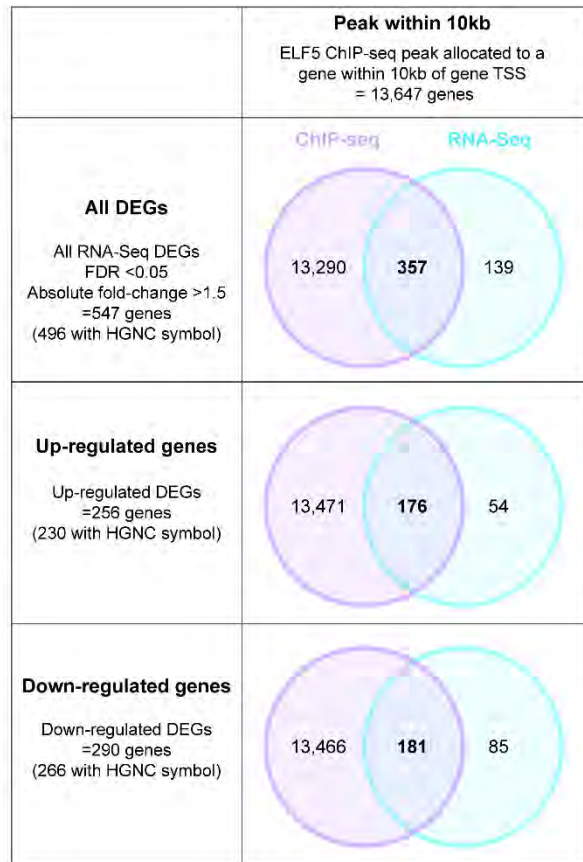
**Figure 4.12: Integration of ELF5 RNA-seq and ChIP-seq to identify direct ELF5 target genes**

**(next page)**

(A) Overlap of ChIP-seq peak-associated genes and differentially expressed genes (DEGs) identified in the MCF7-ELF5-V5 RNA-seq experiment (FDR <0.05 and absolute fold change >1.5). ChIP-seq peak-associated gene lists were generated by allocating each peak to the closest gene and filtering to include only those genes with a peak within 10kb of the transcriptional start site (TSS) in any direction. The overlapping genes, indicated in bold, were classified as direct ELF5 target genes and are specifically indicated in Additional Tables 4.1 and 4.2. (B-C) MSigDB analysis of up-upregulated (B) and down-regulated (C) direct targets for enrichments in GO Biological Process and Hallmark collection gene sets. The top 10 enriched gene sets for each collection are shown, ranked according to the  $-\log_{10}(\text{FDR})$ . The dotted line indicates an FDR of 0.05. HGNC, Human Genome Organisation (HUGO) Gene Nomenclature Committee.



**A**



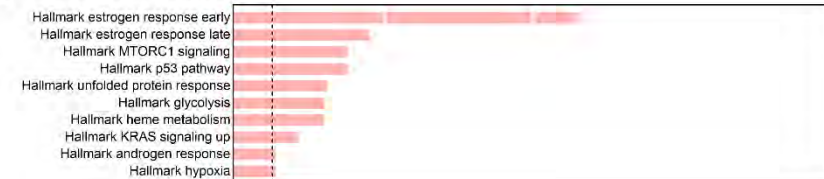
**B**

Direct ELF5 ChIP targets: Up-regulated genes (176)

Top 10 Enriched GO Biological Processes



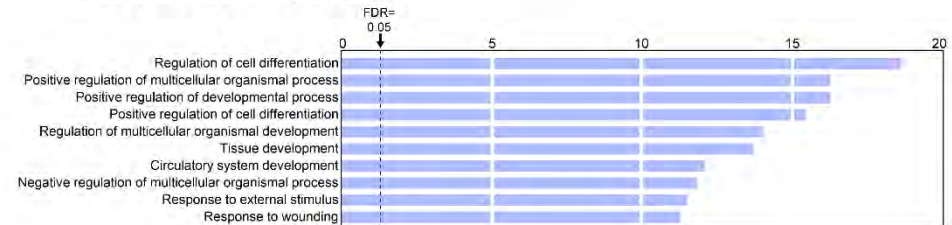
Top 10 Enriched Hallmark Sets



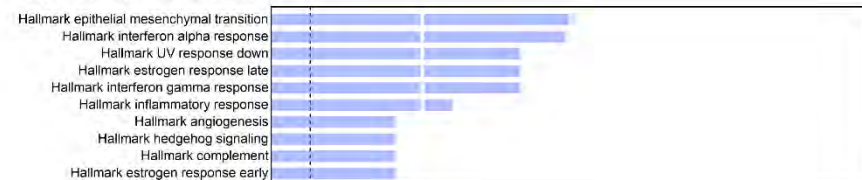
**C**

Direct ELF5 ChIP targets: Down-regulated genes (181)

Top 10 Enriched GO Biological Processes



Top 10 Enriched Hallmark Sets





### **FOXA1 genomic binding sites in the context of low and high ELF5 expression**

Several lines of evidence prompted the investigation of FOXA1 binding patterns in the context of low and high ELF5 expression. Firstly, the Forkhead motif was the second-most enriched motif after the ETS motif at the genomic binding sites of ELF5.

Enrichment for FOXA1 motifs in ELF5 ChIP-seq peaks has also been previously demonstrated in T47D cells (Kalyuga *et al.*, 2012). As FOXA1 is known to function extensively in the regulation of ER binding (Hurtado *et al.*, 2011), it was considered to be a highly likely Forkhead family candidate for binding to these sites and directly or indirectly interacting with ELF5. Secondly, FOXA1/ER redistribution occurs in endocrine resistant cell lines and poor-prognosis breast cancers, with the underlying mechanisms for this molecular event currently unknown (Hurtado *et al.*, 2011; Ross-Innes *et al.*, 2012). Therefore, FOXA1 ChIP-seq was performed in MCF7-pHUSH-ELF5-V5 cells treated with vehicle (referred to as FOXA1-ELF5-low) or doxycycline (FOXA1-ELF5-high) to determine if increased ELF5 expression had any effect on the genomic distribution of FOXA1 binding.

An overview of the FOXA1 ChIP-seq experiments is shown in Figures 4.13 (FOXA1-ELF5-low) and 4.14 (FOXA1-ELF5-high). Two replicates were performed and the consensus peak sets for FOXA1-ELF5-low (12,896 peaks, Figure 4.13A) and FOXA1-ELF5-high (20,703 peaks, Figure 4.14A) were generated by overlapping the peaks identified by two separate peak-callers in both replicates. There is an obvious discrepancy in the number of peaks identified in the two replicates, with significantly fewer peaks identified in replicate 2 in both conditions. Almost all of the peaks identified in replicate 2 were also identified in replicate 1, suggesting that the second experiment has a lower sensitivity than the first. However, this also means that the peaks identified in both replicates, forming the consensus sets, are likely to represent a true and robust subset of the complete set of FOXA1 binding sites.

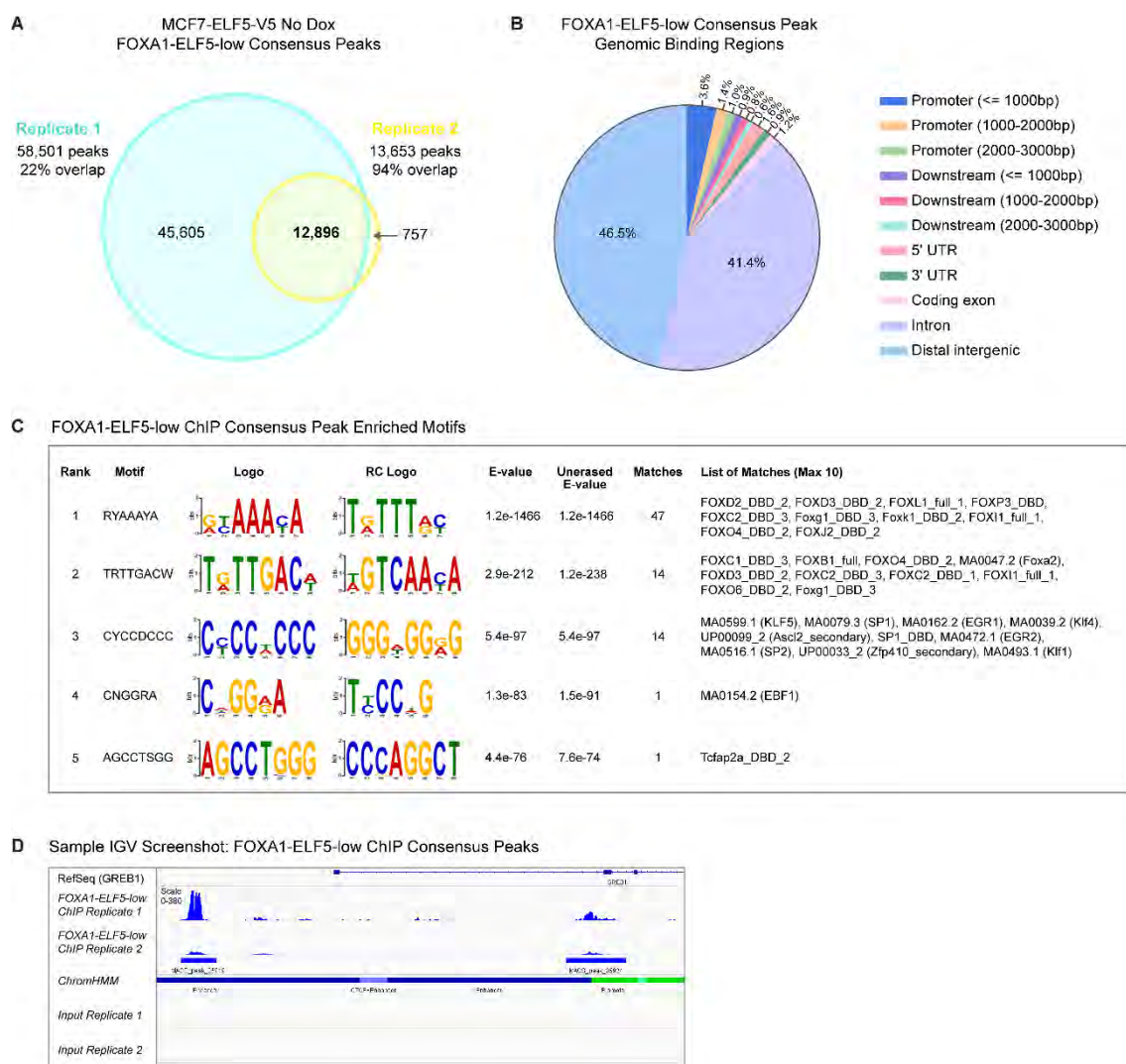
The distribution of the genomic regions of FOXA1 binding were very similar in both ELF5-low and ELF5-high conditions (Figures 4.13B and 4.14B). Approximately 6% of binding sites were located in promoters (<3kb from the TSS); while this is still enriched compared to the genomic background, this is significantly lower than the level of promoter binding seen for ELF5. Almost 90% of FOXA1 binding sites were located in intronic and distal intergenic regions, consistent with the known role of FOXA1 in facilitating the binding of ER to distal enhancers (Carroll *et al.*, 2005).

*De novo* motif analysis demonstrated that a Forkhead motif, matching multiple members of the Forkhead family including FOXA1, was the most enriched motif in the FOXA1 ChIP-seq peaks in both the ELF5-low and ELF5-high experiments (Figures 4.13C and 4.14C). A motif matching an oestrogen response element (ERE) half-site was also enriched in both experiments, ranked 23rd (of 43) by E-value in the FOXA1-ELF5-low experiment and 25th (of 50) in the FOXA1-ELF5-high experiment. The FOXA1-ELF5-high experiment also featured a highly-significant ETS motif, shown in the table in Figure 4.14C (ranked 4th). In the FOXA1-ELF5-low experiment, there was a single enriched ETS motif ranked 19th by E-value. The third-ranked motif in both FOXA1 ChIP experiments, matching to various zinc finger transcription factors such as SP, KLF and EGR family members, was also significantly enriched in the ELF5 ChIP-seq experiment.

Examples of FOXA1 consensus peaks are shown in Figures 4.13D and 4.14D. These peaks are located upstream and within the gene body of the ER-regulated gene growth regulation by oestrogen in breast cancer 1 (*GREB1*). In both the ELF5-low and ELF5-high experiments, a strong peak can be seen in a potential enhancer region approximately 2.5kb upstream from the first *GREB1* TSS. In addition, a second peak is located in the *GREB1* gene body, which may regulate transcription from an alternative promoter. In the FOXA1-ELF5-high experiment, a third enhancer peak is seen approximately 1kb upstream of the TSS. Differences in sensitivity between the two replicates can also be seen in these screenshots, particularly for the FOXA1-ELF5-low experiment in which the peak height in replicate 2 is substantially reduced.

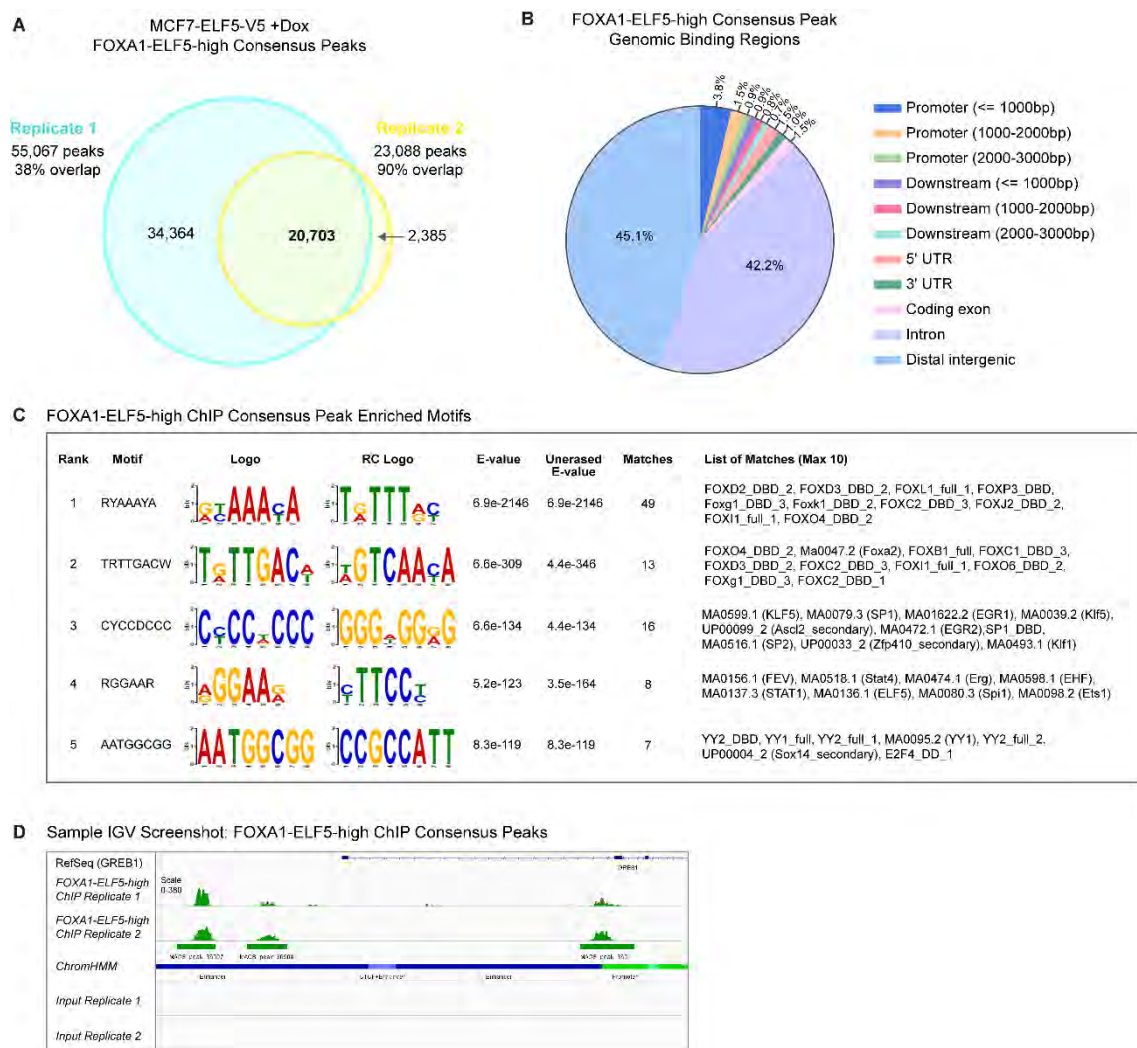
### **Functional analysis of FOXA1- ELF5-low and ELF5-high ChIP-seq consensus peaks**

As for ELF5, functional analyses of the FOXA1 consensus peaks were performed using the Genomic Regions Enrichment of Annotations Tool (GREAT) (McLean *et al.*, 2010). GREAT assigned a total of 23,056 genes to the FOXA1-ELF5-low peaks and 37,157 genes to the larger set of FOXA1-ELF5-high peaks. Figure 4.15A, which shows the data for the FOXA1-ELF5-low experiment, demonstrates that approximately 80% of peaks were associated with 2 genes. The distribution of peaks around the transcriptional start site (TSS) of associated genes is shown for FOXA1-ELF5-low in Figure 4.15B; overall, the distribution for both conditions was extremely similar. In agreement with the genomic binding regions analyses, most peaks were located a substantial distance from their associated genes, with only 6.6% of peaks within 5kb of the gene TSS (6.7% for FOXA1-ELF5-high). Overall, 58.2% of FOXA1-ELF5-low genes



**Figure 4.13: FOXA1-ELF5-low ChIP-seq summary**

(A) Numbers of FOXA1 binding sites identified by both peak-callers (MACS and HOMER) in each replicate, overlapped to produce a consensus set of 12,896 FOXA1-ELF5-low binding sites. FOXA1-ELF5-low is used to refer to FOXA1 ChIP-seq in MCF7-pHUSH-ELF5-V5 cells treated with vehicle and not doxycycline, thereby maintaining ELF5 expression at a low level. (B) Genomic binding regions of the 12,896 FOXA1-ELF5-low consensus peaks, generated using the CEAS tool. (C) *De novo* motif analysis of the 12,896 FOXA1-ELF5-low consensus peaks using DREME and Tomtom. The top 5 motifs identified are represented as forward and reverse logos. The E-value, an indicator of statistical significance, represents the Fisher's exact test p-value multiplied by the number of candidate motifs tested. The best database matches for the identified motifs (up to a maximum of 10) shown on the right. (D) Integrative Genomics Viewer (IGV) screenshots showing examples of FOXA1-ELF5-low binding at two consensus peaks. One peak is upstream of the transcriptional start site of *GREB1* and the other is within the *GREB1* gene body. Chromatin states for MCF7 cells are shown in the ChromHMM track, demonstrating binding of FOXA1 to enhancer and promoter regions in these examples. Minimal signal is seen in the input control samples (bottom two tracks).



**Figure 4.14: FOXA1-ELF5-high ChIP-seq summary**

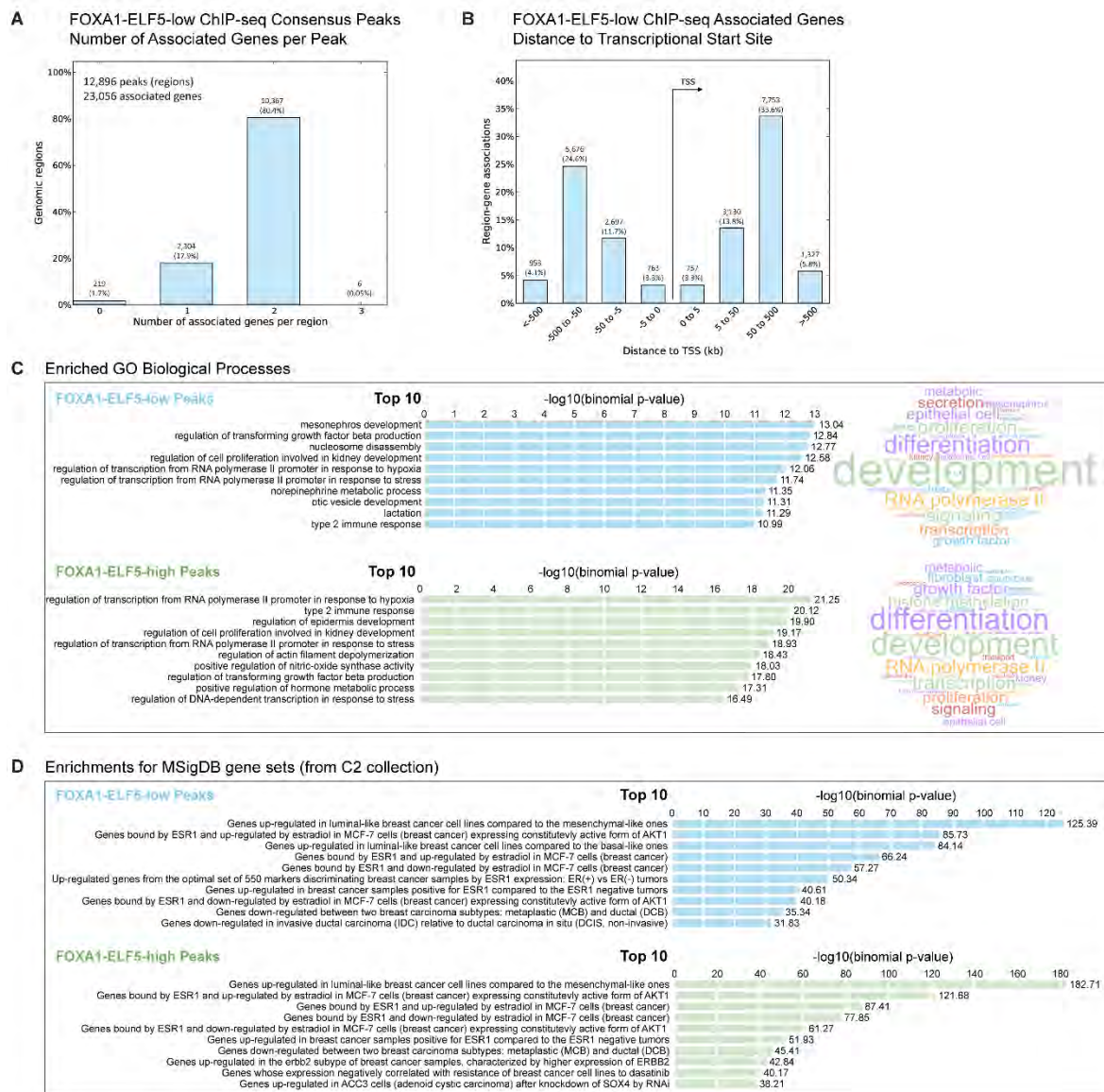
(A) Numbers of FOXA1 binding sites identified by both peak-callers (MACS and HOMER) in each replicate, overlapped to produce a consensus set of 20,703 FOXA1-ELF5-high binding sites. FOXA1-ELF5-high is used to refer to FOXA1 ChIP-seq in MCF7-pHUSH-ELF5-V5 cells treated with doxycycline to induce increased ELF5 expression. (B) Genomic binding regions of the 20,703 FOXA1-ELF5-high consensus peaks, generated using the CEAS tool. (C) *De novo* motif analysis of the 20,703 FOXA1-ELF5-high consensus peaks using DREME and Tomtom. The top 5 motifs identified are represented as forward and reverse logos. The best databases matches for the identified motifs (up to a maximum of 10) are shown on the right. (D) Integrative Genomics Viewer (IGV) screenshots showing examples of FOXA1-ELF5-high binding at three consensus peaks. Two peaks are upstream of the transcriptional start site of *GREB1* and the third is within the *GREB1* gene body. Two of the three peaks are also seen in the FOXA1-ELF5-low ChIP-seq (Figure 4.13D). Chromatin states for MCF7 cells are shown in the ChromHMM track, demonstrating binding of FOXA1 to enhancer and promoter regions in these examples. Minimal signal is seen in the input control samples (bottom two tracks).

were associated with a peak that was 50-500kb from the TSS (58.0% for FOXA1-ELF5-high). As FOXA1 is known to bind distal regulatory regions, it is highly likely that these distal sites are biologically relevant rather than representing false-positive associations.

Extensive analyses of FOXA1 ChIP-seq in MCF7 cells have been previously published (Hurtado *et al.*, 2011; Jozwik *et al.*, 2016; Ross-Innes *et al.*, 2012). Therefore, the functional analyses described here were aimed at: (1) Validating the ChIP-seq performed by confirming the functional relevance of the identified binding sites, and (2) Detecting possible differences in functional enrichments between the ELF5-low and ELF5-high conditions.

In both ELF5-low and ELF5-high conditions, the enriched GO biological processes related to development, differentiation, proliferation, transcription and immune responses, with no clear differences in functional themes emerging. Figure 4.15C shows the top 10 enriched processes for FOXA1-ELF5-low (top) and FOXA1-ELF5-high (bottom) with the key words from the top 100 enriched biological processes represented as a wordcloud. Enrichments for gene sets from the MSigDB C2 collection were also analysed, with Figure 4.15D showing the top 10 sets for each condition. There was a striking enrichment for gene sets related to oestrogen action and breast cancer (19 of 20 sets shown), confirming that the current ChIP-seq experiments did indeed identify binding sites consistent with the known actions of FOXA1 in MCF7 cells. The overlap between the top 100 gene sets for FOXA1-ELF5-low and FOXA1-ELF5-high was 70%, indicating that, as a whole, the identified peaks in both conditions regulate very similar genes. This is not surprising, since there is in fact a large overlap in the binding sites identified in these two conditions (see next section). The 30% difference in the enrichments for the complete peak sets therefore most likely relate to the different numbers of peaks detected in each condition, rather than widespread changes in biological function.





**Figure 4.15: Functional analysis of FOXA1-ELF5-low and FOXA1-ELF5-high ChIP-seq peaks**

Genomic Regions Enrichment of Annotations Tool (GREAT) (A-D) analyses of the FOXA1-ELF5-low (12,896) and FOXA1-ELF5-high (20,703) ChIP-seq consensus peaks.

(A) Number of associated genes per consensus peak for FOXA1-ELF5-low, based on default GREAT settings. There are 23,056 non-unique genes (10,221 unique) associated with the 12,896 peaks, with the majority of peaks associated with 1-2 genes. (B) Associated genes for the FOXA1-ELF5-low peaks grouped by distance of the consensus peak to the transcriptional start site (TSS). A very similar distribution is seen for FOXA1-ELF5-high (data not shown). (C) Top 10 enriched GO Biological Processes (BP) for the FOXA1-ELF5-low (top) and FOXA1-ELF5-high (bottom) consensus peaks, generated by GREAT and ranked by  $-\log_{10}(\text{binomial } p\text{-value})$ . FDR of all sets is  $<0.05$ . The associated wordcloud is based on the top 100 GO BP sets (FDR  $<0.05$ ), with word size proportional to the number of

occurrences (minimum 2). (D) GREAT enriched gene sets from the MSigDB C2 gene set collection (chemical and genomic perturbations subset) for FOXA1-ELF5-low (top) and FOXA1-ELF5-high (bottom) consensus peaks. The top 10 sets are shown, ranked by  $-\log_{10}(\text{binomial } p\text{-value})$ . Descriptive names for gene sets are shown (in contrast to the abbreviated names shown in previous Cytoscape RNA-seq GSEA figures). FDR of all sets is  $<0.05$ .

### **Identification of ELF5-induced changes in FOXA1 binding sites**

The first step in the investigation of ELF5-induced changes in FOXA1 binding was the identification of the peaks that were unique to either FOXA1-ELF5-low or FOXA1-ELF5-high. This was achieved by overlapping the binding sites, defining a common site as one that shared at least one base pair. In addition, the binding sites were overlapped with ELF5 binding sites to provide insights into possible mechanisms of ELF5-induced alterations.

The overlap of the three ChIP-seq experiments is shown in Figure 4.16A and 4.16C, demonstrating two key findings. Firstly, almost all of the FOXA1-ELF5-low binding sites are contained within the larger FOXA1-ELF5-high binding sites, with a total of only 1,319 binding sites present in FOXA1-ELF5-low that are not present in FOXA1-ELF5-high. In contrast, there are a total of 9,126 binding sites present in FOXA1-ELF5-high that are not present in FOXA1-ELF5-low. The second finding is that approximately one-third of all FOXA1 binding sites overlap with an ELF5 binding site.

However, one problem with this direct overlap of peaks is the previously discussed difference in sensitivity between the experimental replicates, particularly in the FOXA1-ELF5-low condition. Therefore, stringent criteria were developed to define a “one-condition-only” peak, as outlined in Figure 4.16B. For a peak to be assigned to the stringent subset of, for example, “FOXA1-ELF5-high-only” peaks, this peak must be present in both replicates of FOXA1-ELF5-high and absent in both replicates of FOXA1-ELF5-low (as shown in column 4). If a peak has been identified in one FOXA1-ELF5-low replicate and not the other (for example, due to low sensitivity of this second replicate), the peak will not be included in the stringent FOXA1-ELF5-high-only subset (as shown in columns 2 and 3).

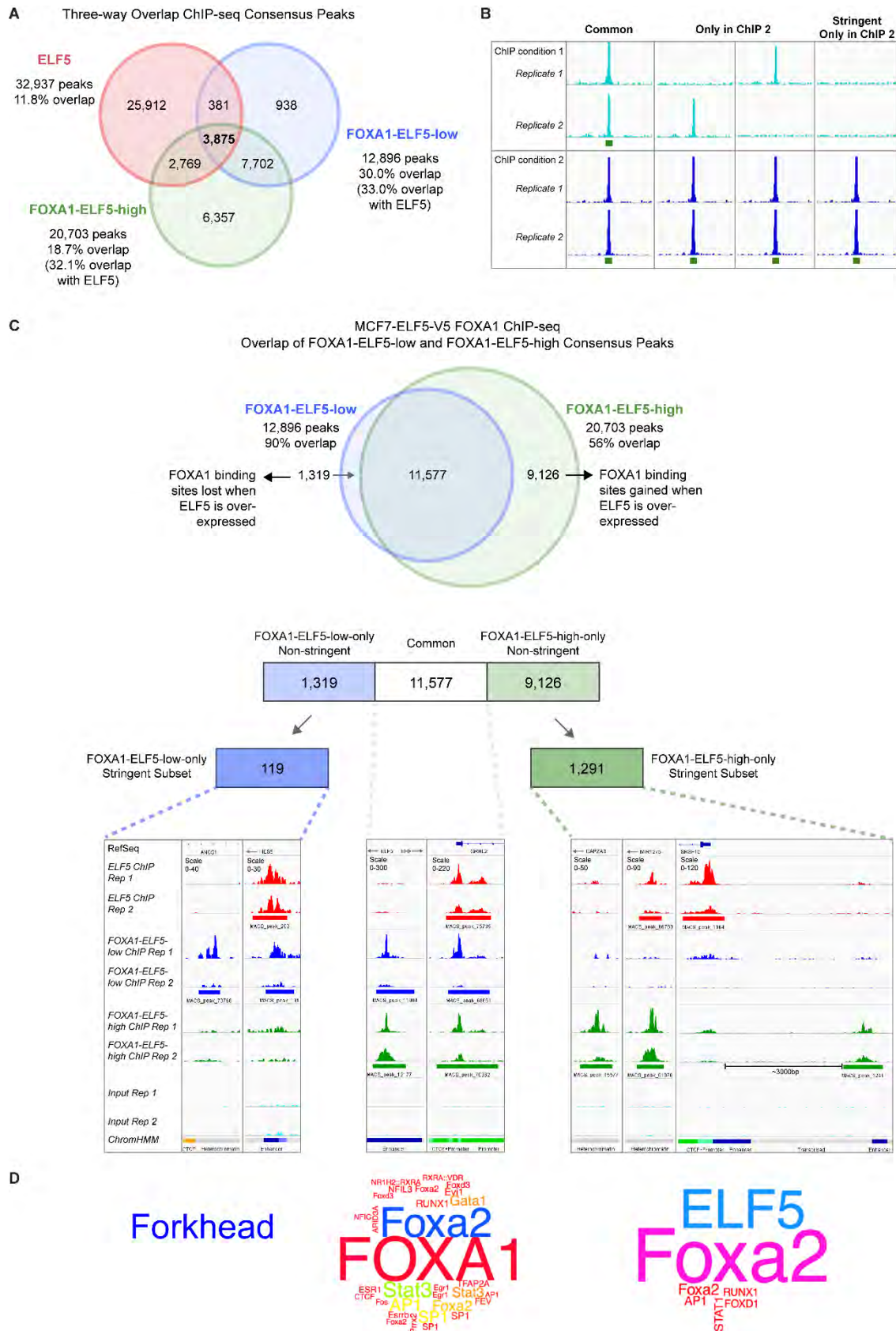
Using these criteria, stringent subsets of 119 FOXA1-ELF5-low-only and 1,291 FOXA1-ELF5-high-only binding sites were defined (Figure 4.16C). These represent high-confidence binding sites that are either lost or gained when ELF5 expression is increased. All future references to FOXA1-ELF5-low-only or FOXA1-ELF5-high-only

binding sites relate to these stringent subsets unless otherwise indicated. On the left, two representative screenshots of FOXA1 binding sites that are only present when ELF5 expression is low (FOXA1-ELF5-low-only) are shown; in one case, the location of the lost peak directly overlaps with that of an ELF5 peak. The screenshots on the right are examples of FOXA1 binding sites that are only present when ELF5 expression is high (FOXA1-ELF5-high-only). In some but not all cases, as for the lost binding sites, the new FOXA1-ELF5-high peaks directly overlap with ELF5 binding sites. A third scenario is the gain of a FOXA1 binding site near, but not directly overlapping with, an ELF5 binding site; in the example shown, a FOXA1 peak is gained in a potential enhancer region approximately 3kb upstream from an ELF5 peak in the serine and arginine rich splicing factor 10 (*SRSF10*) promoter.

*De novo* motif analysis was also performed for the three subsets of peaks (Figure 4.16D). In the common set of peaks, the most enriched motif was a Forkhead FOXA1 motif, with ERE motifs (ESR1) and ETS motifs (FEV) also significantly enriched; this is similar to the results for the complete FOXA1-ELF5-low and FOXA1-ELF5-high sets. In the FOXA1-ELF5-low-only subset, only one motif was identified as significantly enriched, matching to multiple members of the Forkhead family. The small number of peaks, however, is a significant limitation of this analysis. In the FOXA1-ELF5-high-only peaks, there were 9 enriched motifs; the two most significant motifs, by some margin, were a Forkhead motif (E-value =  $8.9\text{e-}96$ ) and an ETS motif (E-value= $1.8\text{e-}51$ ), which had a top database match for ELF5. A second motif analysis tool (MEME) also identified these as the top two motifs and in fact ranked ETS above the Forkhead motif in terms of statistical significance. This strongly suggests that ELF5 binding plays a role in the redistribution of FOXA1 binding to these new sites.

To explore this in more detail, the signal intensity for all three ChIP-seq experiments (and the input DNA as a negative control) was plotted for the peaks in each subset (Figure 4.17). The signal window was centred on the peak summit and extended 5kb in either direction. For comparison purposes, baseline signal levels in “one-condition-only” stringent peak subsets, which compare the ELF5 and FOXA1 peaks, are shown in Figure 4.18; these heatmaps clearly demonstrate increased signal in the relevant “only” ChIP experiment, with a reduced (but not completely absent) signal in the comparison group. This establishes the visual parameters for what constitutes a strong signal, as well as what constitutes a background signal level in the comparison groups. The average binding signal for the individual ChIP-seq experiments in each peak subset is also represented in the plots to the right of the heatmaps.





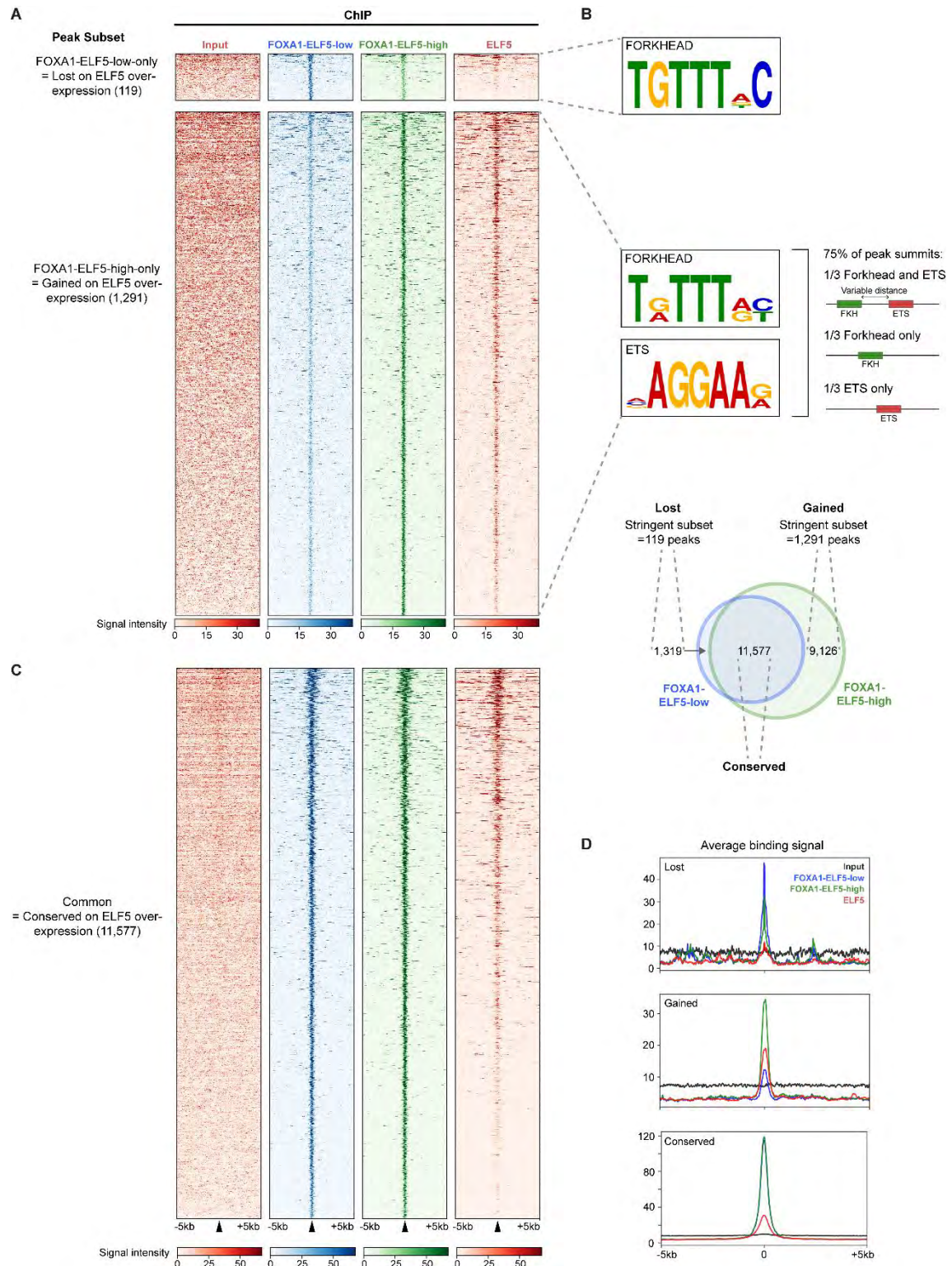
**Figure 4.16: ELF5-driven redistribution of FOXA1 binding**

(A) Overlap of the consensus peaks from the ELF5, FOXA1-ELF5-low and FOXA1-ELF5-

high ChIP-seq experiments. An overlapping peak was defined as a peak sharing at least one base pair. (B) Due to the low number of replicates, stringent criteria were developed to call peaks present only in one condition. To be classed as a condition 2-only peak (column 4), a peak must be called in both replicates of condition 2 (bottom row) and neither replicate of condition 1 (top row). A peak called in one of two replicates in condition 1 is not classed as a stringent condition 2-only peak (columns 2-3). Common peaks have binding in both conditions in both replicates (column 1). (C) Using these criteria, stringent subsets of FOXA1-ELF5-low-only peaks (119) and FOXA1-ELF5-high-only peaks (1,291) were generated. All future references to FOXA1-ELF5-low-only or FOXA1-ELF5-high-only binding sites relate to these stringent subsets unless otherwise indicated. Integrative Genomics Viewer screenshots (separated by vertical lines within boxes) show examples of each category. In the FOXA1-ELF5-low-only subset (119 peaks), the left screenshot shows an example of a FOXA1-ELF5-low-only peak that does not overlap with an ELF5 peak, while the example on the right shows an example that directly overlaps with a region of ELF5 binding. Similarly, in the FOXA1-ELF5-high-only stringent subset (1,291 peaks), FOXA1 peaks may either overlap be ELF5-overlapping (middle) or non-overlapping (left). A third scenario is shown on the right, where a FOXA1 -ELF5-high-only peak occurs in an enhancer region approximately 3kb upstream of an ELF5 promoter peak. (D) *De novo* motif analysis of each subset using DREME and Tomtom. Results are presented as a wordcloud using the highest-ranked database matches and scaled using the DREME E-value (with larger size indicating higher enrichment).

The first horizontal box of the heatmap in Figure 4.17A shows the ChIP-seq signal for the peaks in the FOXA1-ELF5-low-only group (lost on ELF5 over-expression). There is a strong signal in the FOXA1-ELF5-low ChIP-seq, with a reduced signal in the FOXA1-ELF5-high ChIP-seq. The ELF5 signal level, however, is much lower and is comparable to the baseline ELF5 levels seen in Figure 4.18. This indicates that ELF5 binding may not a contributory factor in the loss of these FOXA1 binding sites on ELF5 over-expression. This is consistent with the motif analysis (Figure 4.17B), which, as discussed above for Figure 4.16, identified Forkhead as the only enriched motif.

This contrasts with the findings for the FOXA1-ELF5-high-only subset of peaks. In this case, the strong FOXA1-ELF5-high ChIP-seq signal is accompanied by a strong and consistent ELF5 ChIP-seq signal. The ELF5 ChIP-seq signal is approximately 50% of the “ELF5-only” levels seen in Figure 4.18, as quantified by the average binding signal graphs. This suggests that there is a significant amount of ELF5 binding also occurring at the locations of these gained FOXA1 peaks, consistent with the identification of both Forkhead and ETS motifs in the *de novo* motif analysis.



**Figure 4.17: ChIP-seq signals at genomic locations of redistributed FOXA1 binding**

(A) Signal intensity heatmaps for the FOXA1-ELF5-low-only (top) and FOXA1-ELF5-high-only (bottom) ChIP-seq peak subsets. The rows are centred on the location of the peak summit and span 5kb in both directions. Column 1 is the input DNA (negative control), while columns 2-4 represent the FOXA1-ELF5-low, FOXA1-ELF5-high and ELF5 ChIP-seq experiments respectively. A scale is shown for each ChIP-seq experiment below the

heatmaps, ranging from 0 to approximately 40. (B) Top enriched motifs for each subset identified by DREME/MEME motif analyses and represented as a logo. A summary of the co-localisation of the Forkhead and ETS motifs is shown for the FOXA1-ELF5-high-only subset. (C) Signal intensity heatmap for the set of peaks common to FOXA1-ELF5-low and FOXA1-ELF5-high (i.e. conserved on ELF5 over-expression). Note that the scales for these heatmaps are different to those in panel A, ranging from 0 to approximately 60. (D) Plots showing the average binding signal for each ChIP-seq experiment in the lost (FOXA1-ELF5-low-only), gained (FOXA1-ELF5-high-only) and conserved (common) peak subsets. As for the heatmaps, the average signal is centred on the location of the peak summit and extends 5kb in both directions.

The motifs in this subset of peaks were investigated further using MEME, which provides information on the spatial location of the identified motifs in the set of peak-summit sequences analysed (in this case, 600 of the 1,219) (Figure 4.17B). Approximately 75% (467) of the peak summit regions contained either a Forkhead or ETS motif. Of these 467 sequences, one-third (151) contained a Forkhead motif only, one-third (155) contained an ETS motif only and one-third (159) contained both a Forkhead and an ETS motif. An additional 3 sequences contained a Forkhead motif in combination with a third motif, partially matching to the recognition motif for FOS like 1, AP-1 transcription factor subunit (FOSL1). Even though this was the third-most enriched motif, it occurred only 7 times in total in the 600 sequences analysed, illustrating the highly significant enrichment of the Forkhead and ETS motifs. There were no half or full EREs identified for this subset by either motif analysis tool. In the sequences where the Forkhead and ETS motifs co-occurred, there was no consistent pattern (or “grammar”) in the motif spacing, which ranged from 0-74 base pairs with a median spacing of 20. This suggests that any potential co-operativity between these two DNA-bound transcription factors is unlikely to involve a direct protein-protein interaction and is more likely to involve an indirect mechanism of co-operativity such as assisted loading. Interestingly, approximately 25% of the total peaks analysed (155 of 600) contained an ETS motif only, introducing tethering of FOXA1 by ELF5 as an additional possible mechanism of co-operativity.

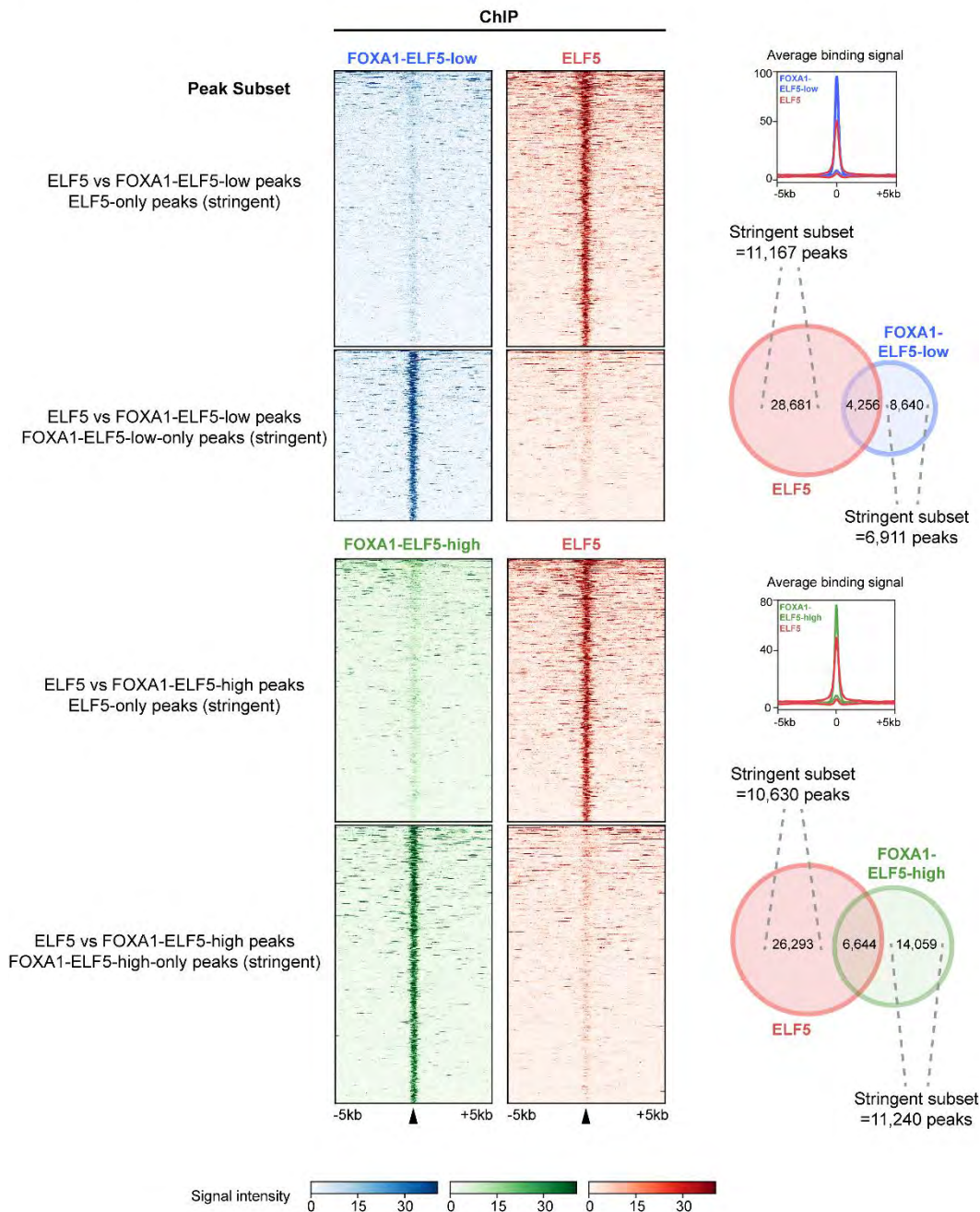
A relatively strong ELF5 signal was also identified in the conserved FOXA1 binding sites (Figure 4.17C, note different scale to panel A). This is consistent with the high level of direct overlap of FOXA1 binding sites with ELF5 binding sites (~33%) seen in the complete datasets (Figure 4.16A).

Finally, the average ChIP-seq binding signal for each peak subset is shown in Figure 4.17D. Interestingly, the FOXA1 sites that are lost or gained on ELF5 overexpression



represent a weaker subset of FOXA1 sites (average signal approx 30-40) when compared to the conserved sites (average signal approximately 120). This suggests that ELF5 may modulate the binding pattern of a subset of weaker FOXA1 binding sites, which may be related to factors such as binding site affinity or accessibility.

Baseline Signal Levels: “one-condition-only” stringent peak subsets



**Figure 4.18: Examples of weak (background) and strong ChIP-seq signals**

Signal intensity heatmaps for stringent “one-condition-only” peak subsets. The subsets are derived from comparing ELF5 with FOXA1-ELF5-low peaks (top), or ELF5 with FOXA1-ELF5-high peaks (bottom). The numbers of peaks in each subset are shown in the Venn

diagrams. The rows are centred on the location of the peak summit and span 5kb in both directions. A scale is shown for each ChIP-seq experiment below the heatmaps, ranging from 0 to approximately 40. An increased signal can be seen in the relevant “only” ChIP experiment, while the signal is reduced but not absent in the comparison group, establishing the background level of binding in the comparison groups. The average binding signal for the individual ChIP-seq experiments in each peak subset is also shown in the plots to the right of the heatmaps.

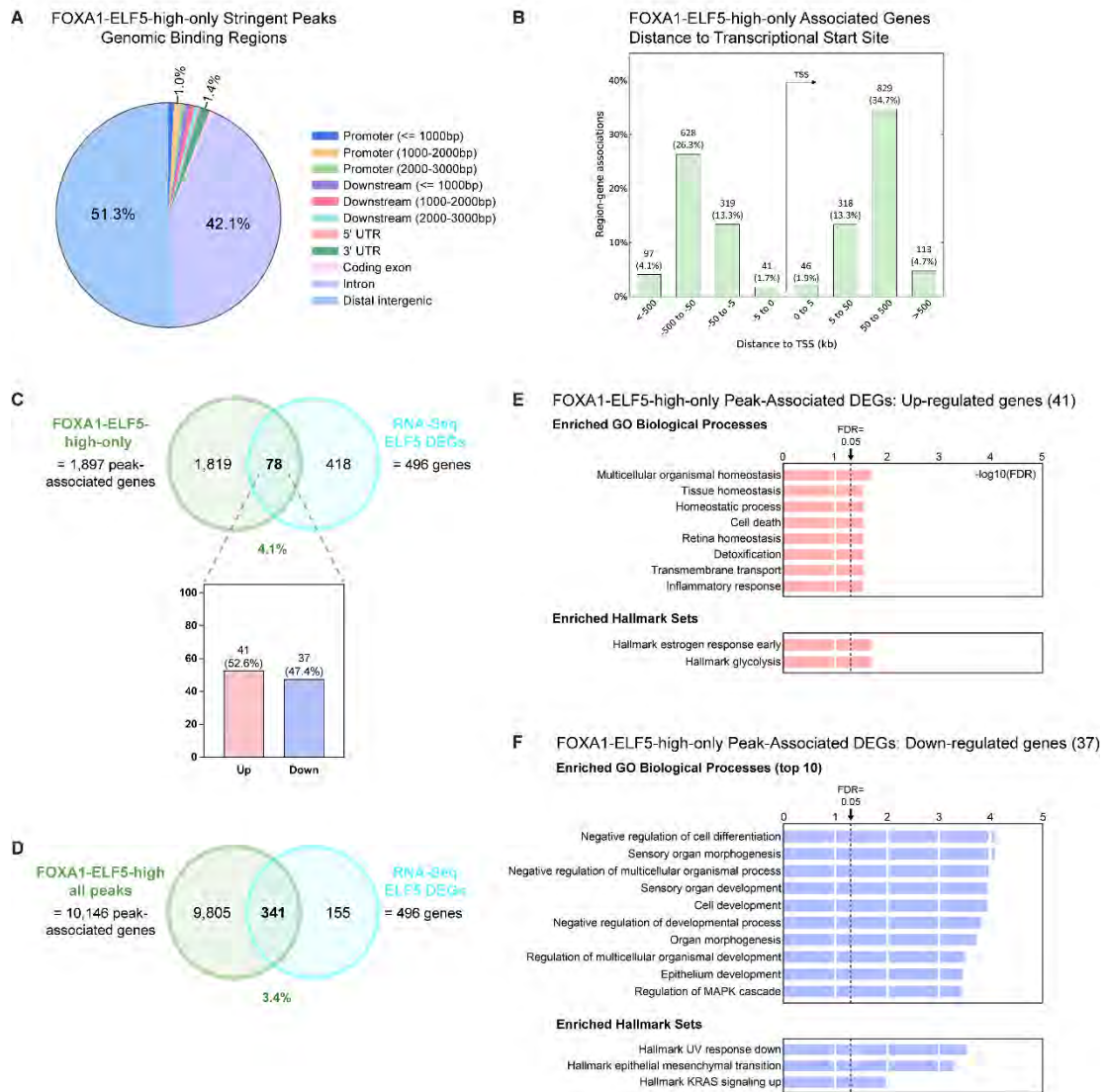
### **Functional analysis of repartitioned FOXA1 binding sites**

The FOXA1 binding sites gained on ELF5 expression (FOXA1-ELF5-high-only) were investigated to determine their potential functional relevance. Firstly, analysis of the genomic binding regions demonstrated that only 2.4% of these sites (or approximately 31 in total) were located in promoter regions less than 3kb from the TSS (Figure 4.19A), which is a reduced proportion compared to the FOXA1-ELF5-high set as a whole. The majority of the gained FOXA1 binding sites were located in distal intergenic and intronic regions (93%).

This distal binding pattern was also demonstrated when GREAT was used to assign peaks to genes (Figure 4.19B). The 1,291 peaks were assigned to a total of 2,391 genes, with the majority of peaks (86%) assigned to 2 genes. Only 3.6% of genes were associated with a peak within 5kb of the gene TSS, with 61.0% of genes associated with a peak located at a distance of 50-500kb.

As gene expression data from the RNA-seq experiment was available, the genes potentially regulated by the binding of FOXA1 to these new sites were examined for changes in expression. Unlike for ELF5, the assignment of FOXA1 target genes based on a proximal distance threshold (for example, 10kb from the TSS) was not considered appropriate, due to the results described above. Therefore, the peak-associated genes from GREAT were used to form a list of potential ELF5-induced FOXA1 regulatory targets. After the removal of duplicates and mapping to official gene symbols, this produced a list of 1,897 potential target genes for the FOXA1-ELF5-high-only binding sites.

This list was then overlapped with the list of differentially expressed genes identified in the RNA-seq experiment (Figure 4.19C). This identified 78 genes that may be in part regulated by ELF5 modulation of FOXA1 binding. These 78 genes represent 4.1% of the total FOXA1-ELF5-high-only gene list and 15.7% of the ELF5 DEGs. Of the 78 genes, approximately equal numbers were up- and down-regulated. As a comparison,



**Figure 4.19: Functional analysis of gained FOXA1 binding sites**

(A) Genomic binding regions of the 1,291 FOXA1-ELF5-high-only peaks, generated using CEAS. All components without a percentage label comprise <1.0% of total. (B) Peak-associated genes, assigned by GREAT, grouped by distance to the TSS. (C) Overlap of GREAT FOXA1-ELF5-high-only peak-associated genes with the differentially expressed genes (DEGs) identified in the MCF7-ELF5-V5 RNA-seq experiment (FDR <0.05 and absolute fold change >1.5). The value below the Venn diagram (4.1%) indicates the percentage of FOXA1-ELF5-high-only peak-associated genes that are also differentially expressed. The column graph shows the proportion of up- and down-regulated genes in the overlapping subset. (D) As for panel C, using the complete set of FOXA1-ELF5-high peak-associated genes (10,146) instead of the FOXA1-ELF5-high-only subset. (E-F) MSigDB analysis of the significantly up- (E) and down-regulated (F) FOXA1-ELF5-high-only peak-associated genes. Enriched GO Biological Process and Hallmark collection gene sets are shown, up to a maximum of 10, ranked according to  $\log_{10}(\text{FDR})$ . Dotted line indicates a FDR of 0.05.

a second GREAT target gene list was generated using the complete set of FOXA1-ELF5-high peaks, which contained 10,146 unique peak-associated genes. Overlap with the RNA-seq DEGs resulted in a higher absolute number of overlapping genes (341), as expected due to the larger size of the input list. However, the proportional overlap was slightly lower (3.4%), suggesting that the FOXA1-ELF5-high-only peak-associated genes may be transcriptionally regulated by the ELF5-induced FOXA1 binding events, although further studies are needed to investigate this possibility.

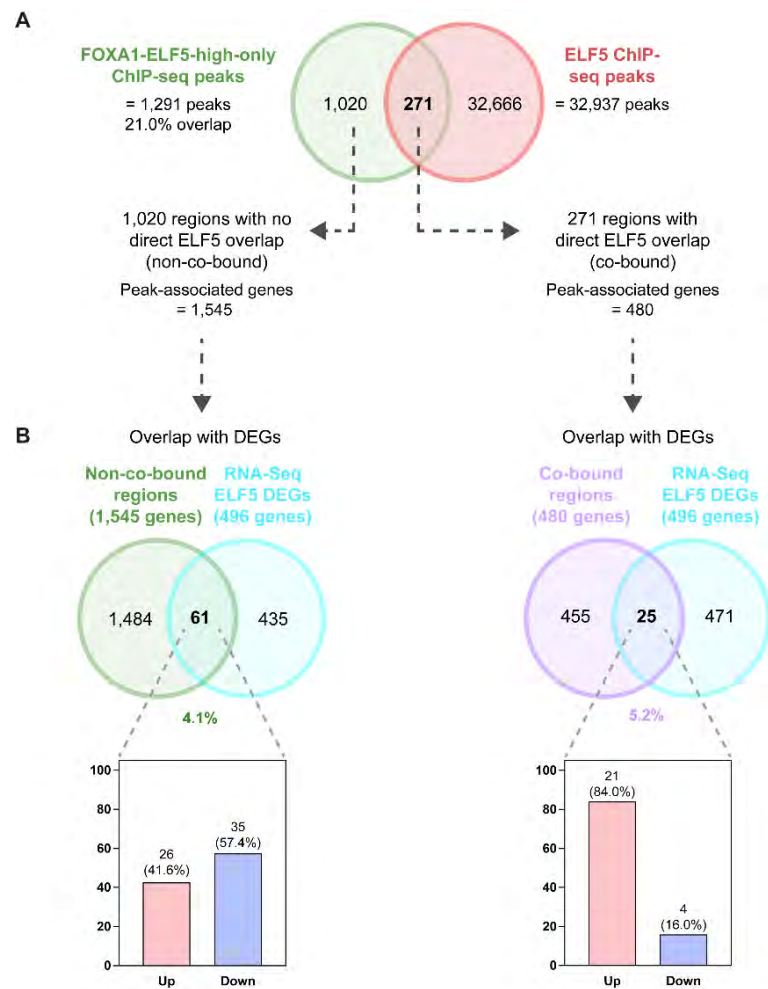
The significantly up- and down-regulated FOXA1-ELF5-high-only target genes were then analysed for enriched gene sets using MSigDB (Figures 4.19E and 4.19F). These up- and down-regulated genes represent a subset of the complete list of ELF5 DEGs that are potentially also regulated by redistributed FOXA1 binding. Several enriched GO biological processes and Hallmark gene sets were identified, although the numbers of overlapping genes were low due to the small input list size. Consistent with previous analyses, there were far more enriched gene sets for the down-regulated genes than for the up-regulated genes (for example, >100 enriched GO biological processes compared to just 8 for the up-regulated genes), suggesting greater functional coherence in the down-regulated genes. Interestingly, there was a distinct lack of immune-related GO BP down-regulated sets, with the vast majority of the top 100 sets related to development and differentiation. There was also no enrichment for the Hallmark oestrogen response-related sets in the down-regulated genes, although the oestrogen response early set was enriched in the up-regulated genes. A possible explanation that unifies these findings is that ELF5 and FOXA1 co-regulate a subset of genes related to promotion of the epithelial identity (whether ER-positive or ER-negative), which represents one of the primary cell-intrinsic functions of both factors.

Next, the direct overlap of FOXA1-ELF5-high-only binding sites and ELF5 binding sites was examined, as the previous motif and signal binding analyses indicated a high degree of co-binding may be present. Of the 1,291 FOXA1-ELF5-high-only binding sites, however, only 271 were found to directly overlap with an ELF5 binding site (Figure 4.20A). This level of overlap (approximately 21% of the FOXA1-ELF5-high-only sites) is in fact slightly lower than the overlap seen for the FOXA1-ELF5-high as a whole (33%). This raises the possibility that simultaneous co-binding may not be the explanation for the results observed.

Peak-associated gene lists for the 271 co-bound and 1,020 non-co-bound regions were again generated using GREAT and compared with the ELF5 RNA-seq DEGs (Figure 4.20B). The overlap for the non-co-bound regions was consistent (4.1% of the peak-



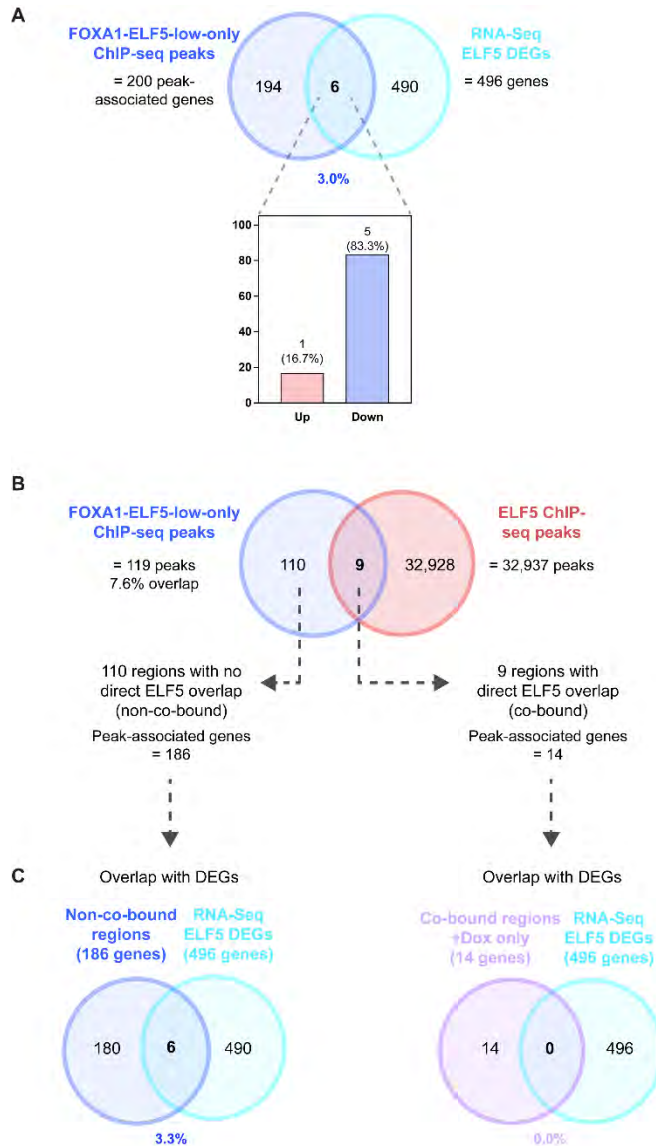
gene list), with a relatively even distribution of up- and down-regulated genes. Interestingly, however, the co-bound regions showed a slightly higher overlap (5.2% of the peak-gene list), with a much larger number of up-regulated genes (21) compared to down-regulated genes (4). Direct overlap of FOXA1 and ELF5 binding may therefore be one mechanism for co-operative up-regulation of a subset of genes.



**Figure 4.20: Analysis of ELF5 co-binding at gained FOXA1 binding sites**

(A) Overlap of the 1,291 FOXA1-ELF5-high-only peaks with the ELF5 consensus peaks, identifying a common set of 271 peaks that share at least one base pair (co-bound regions); the remaining 1,020 peaks were defined as non-co-bound. The numbers of peak-associated genes for each sub-group, assigned by GREAT, are also shown. (B) Overlap of peak-associated genes for each the non-co-bound (left) and co-bound (right) sub-groups with the differentially expressed genes (DEGs) identified in the MCF7-ELF5-V5 RNA-seq experiment (FDR <0.05 and absolute fold change >1.5). The value below each Venn diagram indicates the percentage of peak-associated genes that are also differentially expressed for each sub-group. The column graph shows the proportion of up- and down-regulated genes in the overlapping subset. Note: the number of genes identified in the DEG overlaps (total of 86) is greater than that shown in Figure 4.19 (78) as some genes are associated with multiple peaks that can belong to both sub-groups.

All of the above analyses were also performed for the subset of 119 FOXA1-ELF5-low-only binding sites (that is, the subset of sites lost on ELF5 over-expression). Comparison of the GREAT peak-associated gene list (containing 200 genes) with the ELF5 RNA-seq DEGs demonstrated an overlap of only 6 genes (3.0% of the peak-associated genes), with 1 significant up-regulated gene and 5 significantly down-regulated genes (Figure 4.21A). The low number of peak-associated genes with significant changes in expression precluded any further functional analyses. The FOXA1-ELF5-low-only binding sites were also investigated for direct overlap with ELF5 binding, which revealed a very low direct overlap of only 9 sites (7.6%) (Figure 4.21B). None of the 6 differentially expressed genes were associated with ELF5-overlapping peaks (Figure 4.21C). Overall, this suggests that ELF5-induced loss of FOXA1 binding is not a major regulatory mechanism, although this conclusion is limited by the small number of sites identified in this study.



#### Figure 4.21: Analysis of lost FOXA1 binding sites

*(previous page)*

(A) Overlap of the GREAT FOXA1-ELF5-low-only peak-associated genes with the DEGs identified in the MCF7-ELF5-V5 RNA-seq experiment. The value below the Venn diagram (3.0%) indicates the percentage of FOXA1-ELF5-low-only peak-associated genes that are also differentially expressed. The column graph shows the proportion of up- and down-regulated genes in the overlapping subset. (B) Overlap of the 119 FOXA1-ELF5-low-only peaks with the ELF5 consensus peaks, identifying a common set of 9 peaks that share at least one base pair (co-bound regions); the remaining 110 peaks were defined as non-co-bound. The numbers of peak-associated genes for each sub-group, assigned by GREAT, are also shown. (C) Overlap of peak-associated genes for each sub-group (non-co-bound on the left and co-bound on the right) with the MCF7 RNA-seq DEGs. The value below each Venn diagram indicates the percentage of peak-associated genes that are also differentially expressed for each sub-group.

## Discussion

### Overview

The use of next-generation sequencing technology to examine both gene expression and transcription factor genomic binding sites has provided new insights into the function of ELF5 in a luminal breast cancer context. RNA-sequencing was able to detect more genes at higher fold change levels than microarray technology, leading to the improved identification of several ELF5-regulated transcriptional networks. These included transcriptional signatures of long-term oestrogen deprivation, suppression of the interferon response and MYC-regulated gene expression. In addition, a metabolic signature comprising up-regulation of amino acid transport, protein translation and glycolysis was identified, which is likely to reflect the preparation of the cell for the physiological demands of milk production, and may be mediated by combined ELF5 and MYC transcriptional activity. RNA-sequencing was combined with ELF5 ChIP-seq, which demonstrated that many of these processes are likely to be directly regulated by ELF5 binding to relatively proximal regulatory regions. Finally, increased ELF5 expression was shown to redistribute FOXA1 genomic binding, which may represent a novel mechanism by which the cellular response to oestrogen is modified in both normal development and luminal breast cancer.

### Transcriptional signatures of long-term oestrogen deprivation

A prominent ELF5-driven gene signature emerging from the RNA-seq was the transcriptional response to long-term oestrogen deprivation (LTED) in MCF7 cells (Figure 4.8). LTED is frequently used to model the acquisition of resistance to aromatase inhibitors (AIs), which are an important tool in the treatment of ER-positive breast cancer in post-menopausal patients. AIs prevent the conversion of androgens to oestrogens in peripheral tissues (and breast cancer cells), which represent the primary source of circulating oestrogens following menopause. AIs contrast with other types of endocrine therapies, which directly interact with ER and prevent the binding of endogenous ligand, leading to modulation or down-regulation of ER activity (selective oestrogen receptor modulators and down-regulators) (reviewed in Patani and Martin, 2014).

The adaptation to oestrogen deprivation has been shown to involve multiple molecular mechanisms, primarily invoking an interplay between ER and growth factor signalling pathways. Following an initial period of growth inhibition, a number of models have

demonstrated an acquired hypersensitivity to oestrogen, whereby the tumour cells can proliferate in response to extremely low levels of oestrogen (Johnston *et al.*, 2003; Santen *et al.*, 2004). This may involve an up-regulation of rapid, non-genomic ER signalling events (initiated by membrane-bound ER) and/or genomic ER signalling events. In both cases, the activation of signalling cascades involving receptor tyrosine kinases (for example, ERBB2 and insulin like growth factor receptor 1, IGFR1) and the activation of downstream kinases such as MAPK and AKT are essential, resulting in the phosphorylation and activation of ER in the absence of ligand (reviewed in Nicholson *et al.*, 2004). In most cases, ER expression is retained and is also important for the LTED-adapted phenotype, as treatment of LTED cells with fulvestrant to down-regulate ER activity and protein levels results in a decrease in cell growth *in vitro* (reviewed in Nicholson *et al.*, 2009). Other studies, however, have argued that the genomic actions of ER may not be important in LTED adaptation and that the adaptive gene expression changes are more likely to be regulated by non-ER transcription factors; this may be supported by the relatively poor clinical response rate (around 7%) to fulvestrant in treatment of relapse following AI treatment (Aguilar *et al.*, 2010; Chia *et al.*, 2008; Johnston *et al.*).

Over time (around one year), some models demonstrate progression to an oestrogen independent state, in which oestrogen stimulation has no additional impact on proliferation. This has been hypothesised to result from oestrogen “super-sensitivity” in combination with an increased level of alternate growth factor-regulated pathways that can support maximal cell growth without the requirement for oestrogen (Chan *et al.*, 2002; Nicholson *et al.*, 2004). Adaptation to LTED can also occur via mechanisms that do not involve oestrogen hypersensitivity, for example, in the MCF7-X model involving both oestrogen and exogenous growth factor deprivation (Staka *et al.*, 2005). In addition, up-regulation of growth factor pathways and ligand stimulation can promote cell proliferation in an ER-independent manner (i.e. in the absence of any detectable ER protein expression), indicating that extreme levels of growth factor signalling can promote an entirely oestrogen-insensitive phenotype (reviewed in Nicholson *et al.*, 2004). The mechanisms by which breast cancer cells can adapt to LTED, therefore, are highly complex and are influenced by multiple factors, including cell-extrinsic factors (for example, the availability of growth factors) and cell-intrinsic factors (for example, the expression of ER and other transcription factors).

After just 48 hours of increased ELF5 expression, MCF7 cells showed gene expression changes consistent with an LTED-adapted phenotype. The LTED-altered genes also

regulated by ELF5 included several genes described to function in oestrogen-independent growth in additional studies. One example is the carcinoembryonic antigen related cell adhesion molecule 6 (*CEACAM6*, log2fc 0.87, FDR 0.15), which has been shown to be up-regulated in an independent model of LTED as well as a model of tamoxifen resistance; furthermore, increased *CEACAM6* expression was shown to be significantly associated with relapse in tamoxifen-treated breast cancer (Lewis-Wambi *et al.*, 2008; Maraqa *et al.*, 2008). There were also two ELF5 peaks located approximately 2k and 4kb upstream of the *CEACAM6* TSS, indicating that this may be a true ELF5 transcriptional target despite the modest expression change and FDR.

Insulin like growth factor binding protein 5 (*IGFBP5*) was also significantly up-regulated by ELF5 (log2fc 1.50, FDR 1.0e-6), as well as by LTED in this study. The insulin like growth factor pathway is an important mechanism of LTED adaptation, however the exact role of *IGFBP5* in this process is currently unclear. *IGFBP5*, which binds insulin-like growth factors (IGFs), has been reported to both potentiate and inhibit (by sequestration of IGFs) the IGF signalling pathway; in addition, *IGFBP5* has a number of IGF-independent effects, which may be related to its ability to regulate transcription (Bach, 2015; Xu *et al.*, 2004; Zhao *et al.*, 2006b). In the normal mammary gland, *IGFBP5* is highly expressed during involution, while exogenous over-expression in pregnancy impairs alveolar development and subsequent milk production; these effects have been proposed to result from the inhibition of IGF activity by *IGFBP5* (Marshman *et al.*, 2003; Tonner *et al.*, 2002). The up-regulation of *IGFBP5* by ELF5, which specifies the alveolar lineage, therefore appears counter-intuitive. Importantly, however, the overall expression of *IGFBP5* induced by ELF5 in the RNA-seq experiment remained relatively low at around 6 transcripts per million (TPM); to put this in perspective, this is around one-sixth of the level of ER mRNA expression in MCF7 cells. Modest *IGFBP5* induction by ELF5 during pregnancy may act as part of a regulatory negative feedback loop, insufficient to prevent alveolar development, with further rises in *IGFBP5* occurring during involution to promote apoptosis and remodelling.

An additional role of *IGFBP5* identified in breast cancer cells is the inhibition of non-genomic and genomic ER signalling in an IGF-independent manner, resulting in decreased cell growth (Hermani *et al.*, 2013). In clinical breast cancer samples, higher *IGFBP5* expression has been associated with poorer relapse-free and overall survival (Becker *et al.*, 2012), although another study found the opposite relationship (Ahn *et*

*al.*, 2010). Adding further confusion, a second LTED model demonstrated significant down-regulation of *IGFBP5* in MCF7-LTED cells (Aguilar *et al.*, 2010). There is clearly much to learn about the function of IGFBP5 in breast cancer, particularly in relation to its transcriptional regulation by ELF5 and other factors and its ability to modulate ER activity.

Examination of ELF5-induced expression changes in other members of the IGF signalling pathway revealed non-significant up-regulation of insulin like growth factor 2 (*IGF2*, log2fc 1.63, FDR 0.15) and insulin like growth factor 2 receptor (*IGF2R*, log2fc 0.27, FDR 0.13), while expression of *IGF1R* was unchanged and *IGF1* was not detected. The expression of *IGFBP6*, which primarily acts to inhibit IGF2 action, was significantly down-regulated. In the gene set enrichment analysis, a set of genes up-regulated in response to IGF1 and IGF2 in MCF7 cells was significantly enriched in a positive direction by ELF5 (Pacher Targets of IGF1 and IGF2 Up). Combined, these effects hint at a potential role for ELF5 in up-regulation of the IGF pathway, with an unclear role for IGFBP5, although in the short-term setting these effects appear to be small.

Overall, however, this and previous studies indicate that short-term ELF5 expression is associated with reduced proliferation and a decrease in several growth factor-related pathways in luminal breast cancer cells, including those mediated by epidermal growth factor (EGF), fibroblast growth factors (FGFs) and platelet-derived growth factor (PDGF). The IGF pathway represents a possible exception to this, although any up-regulation of this pathway is insufficient to overcome the acute anti-proliferative effects of ELF5 over-expression *in vitro*. ELF5 ChIP-seq peaks were also strongly associated with target genes up-regulated on growth factor stimulation in the Enrichr analysis of the top 4,000 ChIP-seq (Figure 4.11G), suggesting that ELF5 may directly suppress these signalling pathways. However, of the 1,655 Enrichr-assigned ELF5 peak-associated genes, only 49 were associated with a significant (FDR<0.05) change in gene expression with an absolute fold change of at least 1.5. This indicates that many of the ELF5 peaks included in this analysis (assuming accurate target gene allocation) are not regulating gene expression changes. One possibility is that ELF5 is “bookmarking” sites for regulation in response to future developmental signals, such as growth factors or hormones; this may allow ELF5 to stimulate growth in a context-dependent manner (Wang *et al.*, 2015a). Cells with high ELF5, for example *in vitro* tamoxifen-resistant MCF7 cells (TAMRs), can proliferate and in fact are dependent on ELF5 for continued growth (Kalyuga *et al.*, 2012). A significant proportion of basal-like

breast cancers, which are characteristically highly proliferative, also express high ELF5 (as shown in Chapter 3), with knockdown of ELF5 *in vitro* reducing basal-like cell growth (Kalyuga *et al.*, 2012). In T47D cells, a luminal cell line with relatively high endogenous ELF5 expression, ELF5 has been shown to reduce the level of cell cycle arrest induced by treatment with progestins (Hilton *et al.*, 2010). Furthermore, mammary gland deletion of ELF5 *in vivo* inhibits proliferation of a subset of epithelial cells and reduces phosphorylation of ERK, an important MAP kinase in growth factor-mediated signal transduction (Oakes *et al.*, 2008). This evidence points to the existence of mechanisms by which the growth-suppressive effects of ELF5 can be overcome and, going further, by which ELF5 can even promote oestrogen-independent cell growth in specific contexts. These mechanisms are likely to be essential for the proliferation of ELF5-expressing luminal progenitor cells during pregnancy and can be co-opted in cancer cells to promote growth of an oestrogen-independent phenotype. In some ways, this process may be similar to the adaptation to LTED, which is characterised by an initial period of down-regulation of proliferative genes and growth inhibition, followed by a robust reactivation of these genes leading to renewed growth (Martin *et al.*, 2011). Studies in which low-ELF5 cells are adapted to high increased ELF5 expression over an extended time period may provide further insights into these mechanisms.

Another ELF5 and LTED up-regulated gene is prolactin-induced protein (*PIP*, log2fc 3.29, FDR 0.004) a small glycoprotein synthesised by various glands including the salivary, lacrimal, mammary and prostate glands. The expression of PIP is regulated by hormones such as prolactin, testosterone, growth hormone and glucocorticoids, however the functions of PIP are largely unknown (Baniwal *et al.*, 2014). Recently, knockdown of PIP in the luminal breast cancer cell line T47D was shown to inhibit the activation of several receptor tyrosine kinases and their downstream effectors including AKT and ERK; this resulted in decreased proliferation of both tamoxifen-sensitive and tamoxifen-resistant derivatives without affecting genomic ER activity. PIP has therefore been proposed to be a regulator of diverse growth-promoting signalling pathways in breast cancer cells. This includes those initiated by non-genomic ER as well as by other growth factors, although the exact mechanisms by which this is achieved are unclear (Baniwal *et al.*, 2014). Interestingly, of the genes shown to be positively regulated by PIP (total of 293), there was an overlap of 37 genes also induced by ELF5 (FDR <0.05 with no fold change threshold). These 37 genes included a number of solute carrier family members (6) and metabolic enzymes, producing functional enrichments for amino acid transport and metabolic processes. Other PIP and ELF5



up-regulated genes included *IGFBP5* and *VTCN1* (shown to be direct ELF5 ChIP targets), *MYC*, and the ETS factors *SPDEF* and *EHF*. However, 43 of the up-regulated PIP genes were in fact down-regulated by ELF5, with a strong enrichment for DNA replication and cell cycle-related genes, including minichromosome maintenance complex family members 2-7, DNA polymerase alpha subunit B (*POLA2*) and cyclin-E2 (*CCNE2*). PIP expression, possibly in combination with *MYC* up-regulation (discussed further below), could therefore represent a longer-term mechanism facilitating the growth-promoting effects of ELF5.

### **The interferon response**

Interestingly, examination of the genes down-regulated by LTED adaptation as well as by ELF5 revealed a striking number of genes related to the interferon response. Of the top 20 down-regulated genes in the LTED-down set (ranked by ELF5-induced fold change), 12 were members of the Reactome Interferon Signalling pathway, with an additional 5 identified as interferon-induced genes. These genes included interferon regulatory factors 7 and 9 (*IRF7* and *IRF9*), all four members of the 2'5'-oligoadenylate synthetase family (*OAS1-3* and *OASL*), interferon induced protein with tetratricopeptide repeats 1 (*IFIT1*), interferon induced transmembrane protein 1 (*IFITM1*) and interferon alpha inducible protein 27 (*IFI27*). All of the 17 interferon-related genes were significantly down-regulated in the RNA-seq, with the exception of *IFI44* (FDR = 0.074). The interferon response was also shown to be significantly down-regulated in the gene set enrichment analyses (Figures 4.2 and 4.7B). Furthermore, functional analysis of the ELF5 ChIP-seq peaks demonstrated significant enrichment for interferon-related processes, reinforcing the direct transcriptional effect of ELF5 on this pathway (Figure 4.11C). A number of interferon response genes, including the important regulators *IRF7*, *IRF9* and signal transducer and activator of transcription 1 (*STAT1*), contained an ELF5 ChIP-seq peak in their promoter.

There are three types of interferons (IFNs), type I (including IFN $\alpha$ , IFN $\beta$ , IFN $\epsilon$ , IFN $\kappa$ , IFN $\omega$ ), type II (only IFN $\gamma$ ) and type III (IFN $\lambda$  subtypes), which signal through binding to cell surface receptors. Many cell types, including cancer cells, can express type I and III interferons, while the expression of type II (IFN $\gamma$ ) is mainly restricted to T cells and natural killer cells. However, all types of interferons are capable of regulating fundamental processes in cancer cells, either directly (by binding to receptors expressed by the cancer cell) or indirectly (by regulating the activity of associated immune cells) (reviewed in Parker *et al.*, 2016).

The direct (cell-intrinsic) effects of interferons are primarily tumour-suppressive and include reduced proliferation and increased apoptosis (reviewed in Parker *et al.*, 2016). The RNA-seq data indicate that MCF7-ELF5-V5 cells express very low levels of the IFN $\beta$  (*IFNB1*) and IFN $\lambda$  (*IFNG*) genes, with no expression of other interferons, although they do express all interferon receptor subunits. Overall, the baseline level of interferon signalling in the cultured cells is likely to be low, due to the minimal expression of autocrine interferon and the lack of interferon-expressing immune or stromal cells; the concentration of cytokines in the serum-supplemented growth medium is unknown. However, down-regulation of interferon signalling may contribute to ELF5-mediated tumour progression in the *in vivo* context.

In breast cancer cells, the cell-intrinsic effects of interferon signalling may also influence the response to oestrogen and anti-oestrogens, as suggested by the LTED gene expression changes above. An early indication of the relationship between interferon and oestrogen signalling was the observation that interferon could increase ER expression in breast cancer cells (van den Berg *et al.*, 1987). *In vitro*, the interferon-induced protein interferon regulatory factor 1 (IRF1) is essential for apoptosis of mammary cells in response to tamoxifen or fulvestrant treatment and down-regulation of IRF1 has been associated with endocrine resistance and long-term oestrogen deprivation (Aguilar *et al.*, 2010; Bouker *et al.*, 2004; Bowie *et al.*, 2004; Schwartz *et al.*, 2011). Although *IRF1* expression is not significantly altered in the ELF5 RNA-seq, the expression of beclin 1 (*BECN1*), a negative regulator of IRF1 involved in the balance between autophagy and apoptosis (Schwartz-Roberts *et al.*, 2015), is up-regulated by ELF5 (log2fc 0.34, FDR 0.011). The IRF1-BECN1 feedback loop could therefore be a possible mechanism for ELF5-induced resistance to anti-oestrogen-induced apoptosis in tamoxifen-resistant MCF7 cells (Kalyuga *et al.*, 2012).

Two other IRFs (*IRF7* and *IRF9*) are also significantly down-regulated by ELF5, as well as by LTED in the oncogenic signatures gene set. Down-regulation of genes with IRF motifs in their promoters (for example, *IFI27*, *IFIT1*, *OAS1* and *STAT1*) has also been demonstrated in an independent MCF7 LTED study (Aguilar *et al.*, 2010). However, up-regulation of interferon response genes (including *IRF7*, *IRF9* and *IFITM1*) has also been seen in LTED models (Choi *et al.*, 2015) and a similar up-regulation occurs in a mouse model of tamoxifen resistance (Dabydeen *et al.*, 2015). These conflicting results indicate that there is a need for further studies exploring the specific roles of these interferon pathway proteins in modulation of the endocrine response.

Similarly, conflicting data exist for the interferon induced transmembrane protein 1

(*IFITM1*), which was down-regulated by LTED in the oncogenic signatures gene set, as well as by ELF5 expression. However, in other studies *IFITM1* has been shown to be constitutively over-expressed in LTED cells, promoting the growth of LTED cells *in vitro* and correlating with recurrence following AI treatment in a clinical cohort (Choi *et al.*, 2015; Lui *et al.*, 2017). *IFITM1* has also been associated with increased invasion, metastasis and poor prognosis in several other cancer types, including gastric cancer (Lee *et al.*, 2012), colorectal cancer (Yu *et al.*, 2015), ovarian cancer (Kim *et al.*, 2014), lung cancer (Jin *et al.*, 2017), head and neck cancer (Hatano *et al.*, 2008), and glioma (Yu *et al.*, 2011). Overall, this suggests unique functions of *IFITM1* in the progression of multiple cancer types that contrasts with the primarily tumour-suppressive functions of the interferon pathway. The mechanisms by which this occurs have not been fully characterised, although several studies have implicated an increase in epithelial-to-mesenchymal transition and increased expression of matrix metalloproteinases (Hatano *et al.*, 2008; Lui *et al.*, 2017; Sari *et al.*, 2016); in this context, the down-regulation of *IFITM1* by ELF5 is consistent with the known role of ELF5 in EMT inhibition. These studies demonstrate that the cell-intrinsic effects of the interferon pathway are not exclusively tumour suppressive, as some interferon-induced genes can have tumour-promoting effects.

In the *in vivo* context, cross-talk between tumour cells and immune cells represents an important mechanism of interferon anti-tumour activity. A recent study comparing primary mammary tumours with bone metastases in mice (using 4T1.2 cells) demonstrated that down-regulation of a set of IRF7-regulated genes was crucial for the development of metastatic disease (Bidwell *et al.*, 2012). In the study by Bidwell *et al.*, expression of the interferon-regulated transcription factor IRF7, as well as a set of more than 200 IRF7 target genes, was shown to be down-regulated in bone metastases compared to the primary tumours; conversely, over-expression of IRF7 in the tumour cells decreased the metastatic burden. Importantly, the metastasis-inhibiting effect of IRF7 was dependent on a competent immune system and was shown to result from an interferon-induced reduction in the number of circulating and tumour-associated myeloid-derived suppressor cells (MDSCs). MDSCs suppress the immune response elicited by cancer cells, facilitating their escape from immune surveillance and subsequent metastasis (Law *et al.*, 2017a). Therefore, IRF7 up-regulation in the cancer cells can establish a positive feedback loop, leading to the up-regulation of interferon response genes (IRGs) and interferon, which in turn inhibits the infiltration of MDSCs. Expression of the IRF7 gene signature in the primary tumour was also associated with reduced bone metastases in a clinical cohort (Bidwell *et al.*, 2012).

Intriguingly, increased ELF5 expression has been recently demonstrated to enhance MDSC infiltration in primary PyMT-driven mammary tumours, resulting in an increase in lung metastases. In a transcriptomic analysis of lung metastases compared to primary tumours, interferon-related gene sets were shown to be significantly down-regulated (Gallego-Ortega *et al.*, 2015). In combination, these findings suggest the ELF5-induced down-regulation of the interferon response, possibly through modulation of key regulators such as IRF7, may be an important mechanism facilitating MDSC infiltration and metastasis of breast cancer cells. Furthermore, therapies stimulating the interferon response may be effective in the treatment of ELF5-expressing tumours.

Due to the largely tumour-suppressive effects of interferons, many clinical trials have examined their use in the treatment of breast cancer, both as a single agent and in combination with other treatments such as anti-oestrogens (see Parker *et al.*, 2016 for an excellent review). Overall, however, the results of these trials have been highly variable, which has been attributed to factors such as small patient cohorts, relatively advanced disease stage, dosage inconsistencies, and lack of prospective randomisation. Recent research suggests that the anti-tumour effects of interferons in solid tumours are primarily due to their effects on the immune system, rather than cancer cell-intrinsic effects, and that they are more effective in targeting residual disease or circulating cancer cells than established metastatic disease. It has therefore been proposed that interferons may be effective as an early therapy in patients at high risk of recurrent disease (Parker *et al.*, 2016). ELF5 could therefore have potential as a biomarker predicting both high risk of recurrence (due to decreased sensitivity to anti-oestrogen treatments and increased metastatic propensity), as well as sensitivity to interferon-based therapy (through revitalisation of interferon signalling and inhibition of MDSC activity). Future investigations into the effects of interferons on ELF5-expressing cells may provide further insight into this therapeutic possibility.

## **ELF5 and MYC**

Another central finding from the ELF5 RNA-seq analysis was an enrichment for a gene expression signature consistent with MYC overexpression (Figures 4.8B and 4.8D). The MYC proto-oncogene is a basic helix-loop-helix transcription factor that regulates many cellular processes, including proliferation, cell growth (an increase in cell size), protein synthesis, metabolism, apoptosis, differentiation and angiogenesis (Meyer and Penn, 2008). MYC is also frequently over-expressed in many cancer types, with amplification and/or over-expression occurring in 30-50% of breast cancers (McNeil *et al.*, 2006).

In addition to MYC target genes, *MYC* itself was significantly up-regulated by ELF5 (log2fc 0.52, FDR 2.2e-4), with a broad ELF5 ChIP-seq peak present in the *MYC* promoter and gene body. A second ELF5 peak was located approximately 65-68 kb upstream of the *MYC* TSS, overlapping with an enhancer region that is known to be bound by ER to induce *MYC* expression (Wang *et al.*, 2011). ELF5 has not been previously demonstrated to regulate *MYC* expression. However, another ETS factor (ETS2) binds to the *MYC* promoter in response to growth factor signalling to up-regulate *MYC* expression in endocrine-resistant breast cancer cells (Al-azawi *et al.*, 2007). ETS2 is known to be phosphorylated by MAP kinase, thereby linking growth factor signalling pathways with ETS-mediated transcriptional regulation. The finding in Chapter 5 of this thesis that ELF5 is phosphorylated provides a possible mechanism linking growth factor signalling pathways with the modulation of ELF5 activity to promote growth in an oestrogen-independent context.

MYC is an important mediator of cancer progression in oestrogen-independent breast cancer. It is highly expressed in basal-like breast cancers and is associated with an increased risk of recurrence in ER-positive breast cancers following endocrine therapy. In all subtypes, increased MYC expression has been shown to correlate with higher grade, as well as poorer disease-specific survival and distant-metastasis-free survival (Alles *et al.*, 2009; Green *et al.*, 2016). MYC and its transcriptional targets are also frequently over-expressed in *in vitro* models of endocrine resistance, with an LTED MYC-regulated gene signature predictive of recurrence following adjuvant tamoxifen (Aguilar *et al.*, 2010; Miller *et al.*, 2011). Furthermore, co-expression of LTED and MYC activation signatures identified from oestrogen-deprived cell lines was associated with a poorer outcome than expression of either signature alone (Miller *et al.*, 2011). Interestingly, despite having high MYC expression, amplification of the *MYC* gene is not significantly correlated with the basal-like subtype, suggesting that other mechanisms of increased expression (for example, altered transcriptional regulation) are likely to be important (Alles *et al.*, 2009). Up-regulation of MYC has been proposed to substitute for many of the functions of ER when it is absent. Over-expression of MYC, for example, can restore proliferation in cells treated with anti-oestrogens. MYC and ER regulate a large number of common genes in ER-positive breast cancer cells and enrichment of this MYC-ER gene set is also seen in ER-negative breast cancers (Alles *et al.*, 2009; Musgrove *et al.*, 2008). The up-regulation of *MYC* by ELF5 may therefore be an important mechanism of ER-independent cancer progression. It may also provide an additional explanation for why some genes are similarly regulated (i.e. regulated in the same direction) by ELF5 and ER. Two examples of MYC/ER up-

regulated genes that are also up-regulated on ELF5 over-expression are *DEPTOR* and *MYB*; these genes may be indirectly up-regulated by MYC activity, although it is noted that both genes also have an ELF5 ChIP-seq peak in their promoter region.

The identification of the enriched MYC oncogenic signature sets in ELF5 RNA-seq indicates that MYC is likely to be transcriptionally active in these cells. The fold change increase of 1.43, however, is fairly modest. Relatively small changes in mRNA expression associated with strong MYC transcriptional signatures have also been previously observed in LTED cell lines as well as clinical cohorts of basal-like breast cancers (Alles *et al.*, 2009; Miller *et al.*, 2011). This suggests that mechanisms in addition to mRNA up-regulation may be functioning to increase MYC activity in the ER-independent setting, such as post-translational modifications (which may affect activity or stability), or alterations in co-factor levels or activity. These mechanisms are yet to be investigated in the context of ELF5-induced MYC action.

The role of MYC in promoting proliferation has been well-characterised (reviewed in Bretones *et al.*, 2015); in contrast, short-term ELF5 over-expression produces a clear down-regulation of the cell cycle, despite up-regulation of MYC. However, as discussed above, breast cancer cells with high ELF5 are able to proliferate in certain contexts (including many basal-like breast cancers), suggesting that a shift occurs whereby this inhibitory effect is dampened. The up-regulation of MYC may be one mechanism by which this shift could occur in cells that have adapted to increased ELF5 expression.

A more detailed investigation of the MYC oncogenic signature sets (based on the top 200 up- and down-regulated genes from Bild *et al.*, 2006) in fact revealed very few cell cycle-related genes. Rather, gene ontology analysis of the sets revealed an enrichment in the up-regulated genes for RNA processing and metabolic pathways, and various pathways in the down-regulated genes, including those relating to signal transduction, cell death, development and metabolism. These reflect some of the many known MYC functions that are independent of proliferation (Meyer and Penn, 2008). Recent studies have suggested an important role for a number of these functions in breast cancer, particularly in the setting of endocrine resistance.

An example is the ability of MYC to regulate cell growth (an increase in cell size) and cell metabolism. MYC regulates a number of cell growth processes, including RNA processing, ribosome biogenesis and protein synthesis. Furthermore, increased expression of a MYC-regulated “cell growth” gene signature is associated with shorter distant metastasis-free survival in patients with ER-positive breast cancer treated with

tamoxifen (Musgrove *et al.*, 2008). A recent study has suggested that the functions of MYC may be subtype-specific, with MYC expression associated with the up-regulation of genes related to protein translation and protein metabolism in ER-positive tumours, and up-regulation of genes related to glucose metabolism in ER-negative tumours (Green *et al.*, 2016). The ability of MYC to globally amplify protein production may be an important mechanism by which ELF5 prepares the mammary alveolar cell for large-scale milk protein production.

Along similar lines, MYC also regulates cellular metabolism, up-regulating glycolysis, lactate production, glutamine utilisation, and amino acid transport and biosynthesis. In particular, cancer cells with MYC over-expression commonly use the amino acid glutamine as an energy source, catabolising glutamine to  $\alpha$ -ketoglutarate, which can then feed into the tricarboxylic acid (TCA) cycle (reviewed in Wahlström and Arsenian Henriksson, 2015). Glutamine transport and catabolism are increased by MYC through up-regulation of glutamine transporters (for example, SLC1A5, SLC1A2, SLC7A1) and the enzyme that converts glutamine to glutamate in the first step of glutaminolysis (glutaminase, GLS); these genes are also up-regulated in the ELF5 RNA-seq. In addition, glutamine plays an important role as a nitrogen donor in the synthesis of nucleotides and non-essential amino acids, and regulates the activation of mTOR, which promotes protein translation (Wise and Thompson, 2010). In LTED models of AI resistance, growth factor signalling leads to ligand-independent up-regulation of MYC by ER and an increased level of glutamine consumption, associated with up-regulation of genes such as SLC1A5. Furthermore, inhibition of glutamine metabolism (for example, through inhibition of SLC1A5 or GLS function) results in decreased proliferation of LTED cells; this is due to the phenomenon of “glutamine addiction”, whereby MYC-induced metabolic reprogramming results in an inability of cells to survive in the absence of exogenous glutamine (Chen *et al.*, 2015; Wise and Thompson, 2010). This provides a therapeutic opportunity to target glutamine metabolism in MYC-, and potentially ELF5-, over-expressing breast cancers.

There is also some evidence to suggest a relationship between ELF5 and MYC in normal mammary development. Interestingly, in the normal virgin mammary epithelium, *Myc* expression is higher in the luminal ER-negative cells than in the ER-positive cells; *Myc* is also highly expressed in the basal/myoepithelial cells and has been proposed to be a key regulator of mammary cell fate (Kendrick *et al.*, 2008). Early studies demonstrated that mammary over-expression of *Myc* during pregnancy impaired lobuloalveolar development and lactation (Andres *et al.*, 1988). Intriguingly, this has

been shown to be due to Myc-induced enhancement of alveolar differentiation during pregnancy (specifically between days 12.5 to 15.5), leading to milk production prior to parturition, milk stasis, and premature involution (Blakely *et al.*, 2005). In some ways, this resembles Elf5 over-expression during early pregnancy, which leads to increased expression of milk proteins by day 12.5 (Oakes *et al.*, 2008). However, Elf5 levels continue to rise during pregnancy and lactation, and forced Elf5 expression during the later stages of pregnancy has no additional effect; in contrast, endogenous Myc expression in the mammary gland has been shown to peak during early-mid pregnancy and subsequently fall (Blakely *et al.*, 2005). The over-expression of Myc during the day 12.5-15.5 window results in increased proliferation of alveolar cells and phosphorylation of Stat5, suggesting a link between Myc and the prolactin signalling pathway. The premature activation of Stat5 also occurs with deletion of negative regulators of the prolactin-Stat5 signalling pathway, for example the suppressor of cytokine signalling 1 (Socs1) knockout mouse (Lindeman *et al.*, 2001); this suggests that Myc over-expression leads to a hyperactivation of the prolactin-stat5 pathway.

Conditional deletion of *Myc* in the luminal alveolar cells from mid-pregnancy also leads to a lactation defect, characterised by a decrease in the volume of milk production and reduced alveolar size. The primary cause was shown to be a decrease in translational efficiency, rather than a direct effect of Myc on milk protein production, associated with reduced expression of ribosomal proteins and RNA, as well as proteins involved in translation and ribosome biogenesis (Stoelzle *et al.*, 2009). This indicates that Myc plays an important role in up-regulation of the biosynthetic capacity of the cell to enable large-scale milk production; importantly, this function of Myc remains important in mid-pregnancy and into lactation, despite the tapering of *Myc* expression that occurs during this period. In subsequent pregnancies (in which *Myc* is deleted from the start of pregnancy), there was a lower level of proliferation of the alveolar cells and milk production was delayed but not blocked; however, the proportion of cells that retain *Myc* expression cannot be definitively determined and it is possible that the proliferative capacity is sustained by this population (Sodir and Evan, 2009). The proper timing and co-ordination of Myc and Elf5 expression, therefore, appear to be essential for both alveolar proliferation and the enormous up-regulation of protein synthesis that must occur during pregnancy and lactation. The findings in this chapter, combined with the above studies, provide an intriguing basis for further exploration of the relationship between Myc and Elf5 in normal mammary development and cancer.



### Single-cell heterogeneity and the dynamics of differentiation

An interesting difference between the ELF5 RNA-seq and microarray experiments was a subset of breast cancer subtype and oestrogen response gene sets that were up-regulated in the RNA-seq and down-regulated in the microarray (Figures 4.5 and 4.7). Further investigation of these sets revealed a dynamic, two-component effect of ELF5 on oestrogen-regulated genes, with strong down-regulation of some genes (for example, *KLK11*, *ACOX2*, *TRIM29*, *AREG*) and up-regulation of others (for example, *AQP3*, *MSMB*, *DEPTOR*, *MUC1*). The process of differentiation probably does not occur in discrete stages (as shown in Figure 4.1, for example) but in a much more dynamic and continuous manner. It is proposed, therefore, that this two-component response reflects the dynamics of this process in combination with the heterogeneity of the cells in the culture.

Breast cancer cell lines do not represent homogeneous populations of cells; rather, they are composed of cells with varying differentiation states. Based on the expression of cell surface markers (EpCAM, CD49f and CD24), four epithelial differentiation states can be detected in cells from the normal mammary gland. These varying differentiation states are also present in cultures of breast cancer cell lines (Keller *et al.*, 2010). MCF7 cells in fact represent one of the most diverse breast cancer cell lines of the 16 investigated in this study, containing cells corresponding to mature luminal cells (EpCAM+/CD49f-, around 30%), luminal progenitor cells (EpCAM+/CD24+,/CD49f+, around 50%), mesenchymal cells (EpCAM-/CD49f+, around 10%) and basal cells (EpCAM+/CD24-/CD49f+, <1%); in addition, around 10% of cells were negative for both EpCAM and CD49f. This demonstrates that there is a wide range of cellular states, even when using just three differentiation markers. Individual cells, therefore, are likely to respond differently to cell fate cues such as an increase in ELF5 expression, depending on the underlying differentiation state. Some cells are likely to readily respond to this cue and shift towards a more oestrogen-insensitive phenotype, while others are likely to be more resistant. Due to the dynamic nature of the differentiation process, individual cells in the culture will also be at different points along this differentiation spectrum at the time of collection. Furthermore, every cell is likely to express different amounts of ELF5 from the pHUSH vector due to the random variation inherent in the creation of a pooled cell line. All of these factors may contribute to a heterogeneous response to increased ELF5 expression.

RNA-sequencing was selected for this study due to the increased dynamic range and sensitivity compared to microarrays. In this respect, there is a clear difference between

the two experiments, with the microarray detecting more differentially expressed genes overall (at an FDR <0.05) but far fewer when fold change filters were applied to minimise false-positives. The RNA-seq experiment also detected a much more even proportion of up- and down-regulated genes than the microarray experiment. It is proposed that the increased sensitivity of the RNA-seq experiment, particularly in the detection of up-regulated genes, enabled more of the innate variability present in the culture to be captured, accentuating the two-component aspect of the ELF5 response. As the field moves towards single-cell transcriptomics, it is likely that the dynamics of differentiation will be able to be captured even more fully.

### **The relationship between ELF5 and FOXA1**

The second part of this chapter began with a study of ELF5 genomic binding sites in MCF7-pHUSH-ELF5-V5 cells using ChIP-seq, with a view to defining the direct transcriptional targets of ELF5 that may modulate endocrine sensitivity. An intriguing finding arising from this was the significant enrichment for Forkhead motifs at the sites of ELF5 binding. A number of Forkhead (FOX) transcription factors have been shown to play a role in breast cancer and endocrine resistance, including FOXA1 (Hurtado *et al.*, 2011; Ross-Innes *et al.*, 2012), FOXM1 (Millour *et al.*, 2010; Sanders *et al.*, 2013), FOXC1 (Han *et al.*, 2017), and FOXO3A (Bullock, 2016). Several factors, however, prompted the decision to focus further investigations on FOXA1. Firstly, FOXA1 acts as a pioneer factor to regulate ER binding in normal mammary epithelial cells and in breast cancer cells (Hurtado *et al.*, 2011); the transcriptional balance between FOXA1/ER and ELF5 is, therefore, a key factor in determining cell fate in the mammary gland. Secondly, ELF5 has been previously shown to directly down-regulate FOXA1 expression in the luminal breast cancer cell line T47D (Kalyuga *et al.*, 2012). ELF5 expression has also been shown to be important in maintenance of the basal-like breast cancer molecular phenotype, while FOXA1 directly represses the basal-like phenotype (Bernardo *et al.*, 2012). Once again, these findings reinforce the antagonistic nature of the FOXA1 and ELF5 relationship in both normal development and cancer. Finally, the genomic distribution of FOXA1/ER binding sites is significantly altered in endocrine resistant cell lines and poor-prognosis breast cancers; however the molecular mechanisms behind this redistribution remain unknown (Ross-Innes *et al.*, 2012). It was therefore hypothesised that ELF5 may modulate the oestrogen sensitivity of luminal breast cancer cells through the alteration of FOXA1 genomic binding. This was investigated by performing FOXA1 ChIP-seq in the context of low and high ELF5 expression.

### **FOXA1 is known to regulate ER-chromatin interactions**

FOXA1 belongs to the Forkhead family of transcription factors, characterised by the presence of a Forkhead (or “winged-helix”) DNA-binding domain. There are three members of the FOXA sub-family (FOXA1-3), which have both unique and overlapping functions in normal development. In organs such as the pancreas, kidney, liver, lung and brain, FOXA2 can, for the most part, compensate for deletion of FOXA1 during embryonic development, although combined deletion results in severe defects (reviewed in Bernardo and Keri, 2012). In contrast, there is a unique requirement for FOXA1 in the development of hormone-responsive tissues such as the prostate and mammary gland. In the mammary gland, FOXA1 deletion results in impaired ductal morphogenesis, similar to the deletion of ER. The development of alveolar secretory cells during pregnancy, however, is unaffected, indicating that terminal differentiation of these cells occurs independently of FOXA1. In fact, alveoli in the FOXA1 heterozygous knockout (FOXA1<sup>+/-</sup>) mammary glands are increased in number compared to wild-type glands in response to hormonal stimulation, suggesting that FOXA1, in contrast to ELF5, may normally act to suppress alveolar differentiation (Bernardo *et al.*, 2010).

The requirement for FOXA1 during development of hormone-responsive tissues is likely to be related to its unique role in the regulation of steroid hormone receptors such as androgen receptor (AR) and ER. FOXA1 also acts to regulate the action of these hormone receptors in the context of breast and prostate cancer. As a pioneer factor, FOXA1 can interact with nucleosomal DNA to initiate a local increase in chromatin accessibility, guiding the genomic binding of ER and AR to co-operatively initiate regulatory events in previously silent chromatin. Approximately 50% of ER binding sites and 70% of AR binding sites in breast and prostate cancer cells are co-occupied by FOXA1 (Robinson and Carroll, 2012). Silencing of FOXA1 in MCF7 breast cancer cells results in a global decrease in chromatin accessibility, associated with reduced ER binding at more than 90% of genomic sites. Conversely, overexpression of FOXA1 in non-mammary cells (for example, U2OS-ER osteosarcoma cells) is sufficient to induce ER binding at breast cancer-specific sites (Hurtado *et al.*, 2011). This suggests that the primary positive determinant of ER binding, and therefore function, is the binding of the pioneer factor FOXA1.

### **ELF5 alters FOXA1 genomic binding**

In order to determine the effect of increased ELF5 expression on FOXA1 genomic binding, FOXA1 ChIP-seq was performed in the context of low and high ELF5

expression. One hypothesis was that ELF5 and FOXA1 might compete for the same binding sites, enabling ELF5 to oppose ER-driven changes in gene expression. However, the results of the FOXA1 ChIP-seq suggested that this was unlikely, as increased ELF5 expression only resulted in the loss of 119 FOXA1 binding sites. In contrast, increased ELF5 expression promoted the gain of 1,291 FOXA1 binding sites. Furthermore, there was a strong enrichment for the ELF5 motif at these new FOXA1 binding sites, indicating that ELF5 may act to direct FOXA1 to new genomic regions.

This is an intriguing finding, as unique FOXA1/ER binding patterns have been associated with endocrine resistance and poor-prognosis breast cancers. Redistribution of FOXA1 binding is seen in tamoxifen-resistant cell lines, which rely on FOXA1 expression for continued growth (Hurtado *et al.*, 2011). More recently, ER ChIP-seq in primary and metastatic tumour samples demonstrated a redistribution of ER binding that was correlated with poor patient outcome. The altered ER binding sites in poor-outcome tumours were associated with significant changes in gene expression, and motif analysis and cell line data strongly supported a FOXA1-mediated mechanism of ER redistribution. Interestingly, the ER binding pattern in MCF7 cells, derived from metastatic cells in a pleural effusion, was more similar to the poor-outcome tumours than the good-outcome tumours; this led to the proposal that these cells represent an intermediate ER-binding profile that requires the acquisition of additional ER binding regions to develop an endocrine-resistant phenotype. All of the metastatic samples retained ER binding, suggesting that reprogrammed ER/FOXA1 binding is important for cancer progression (Ross-Innes *et al.*, 2012). These findings indicate that the role of FOXA1 is altered in poor-outcome tumours, with FOXA1-driven reprogramming of ER binding having important implications for clinical progression and response to treatment.

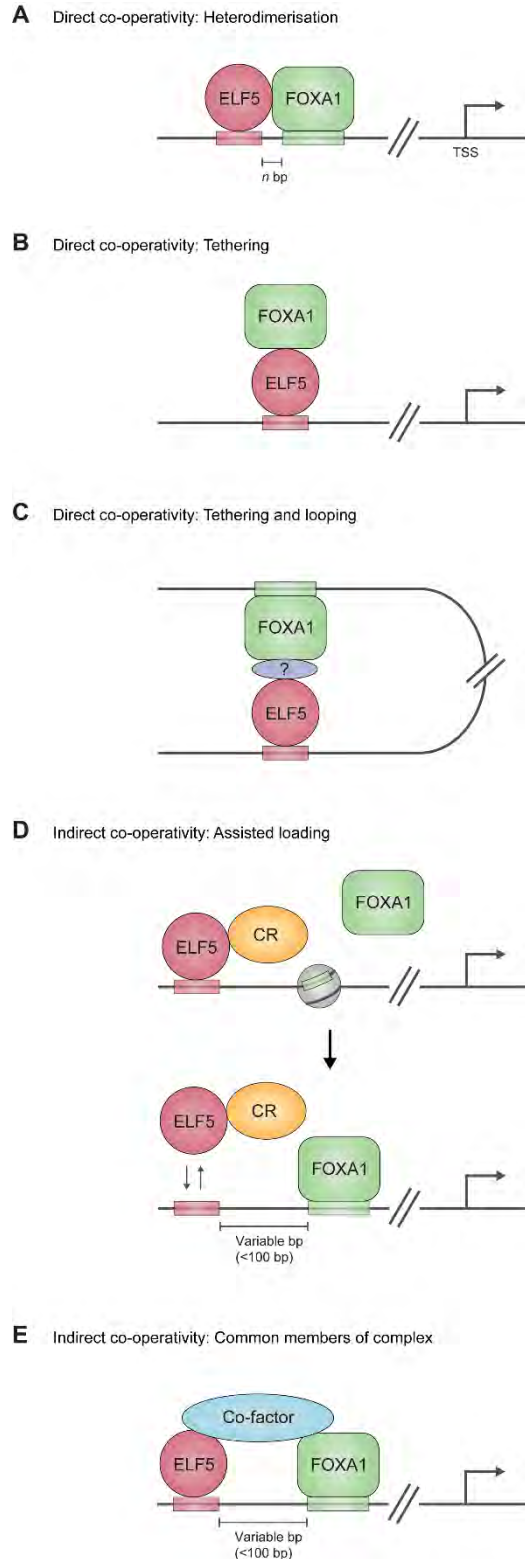
One mechanism that has been shown to reprogram FOXA1 binding is growth factor stimulation, which is also a well-established mechanism of resistance to endocrine therapies (reviewed in Clarke *et al.*, 2015). Growth factor pathways are also a possible mechanism by which ELF5 may promote oestrogen-independent growth (discussed above). Growth factors such as epidermal growth factor (EGF) have been previously shown to promote the ligand-independent binding of ER to distinct genomic sites, with EGF-unique ER binding sites enriched for Forkhead and AP1 transcription factor motifs (Lupien *et al.*, 2010). In a cell line study complementing the ER ChIP-seq clinical study described above, ER was shown to be recruited to approximately 6,000 new binding sites following treatment with a combination of growth factors. Interestingly, more than

50% of these new sites overlapped with FOXA1 binding sites, which were either pre-bound by FOXA1 but not ER (25%) or which gained FOXA1 binding on growth factor treatment (37.5%) (Ross-Innes *et al.*, 2012). This suggests that growth factor stimulation might be one pathway for reprogramming FOXA1 binding in the context of poor-prognosis disease. However, the underlying mechanisms that transduce these growth factor signals to the nucleus and alter FOXA1 genomic binding have not been elucidated; the possibilities include histone modifications, changes in FOXA1 structure and function (for example, through post-translational modifications), or alterations in co-factor levels or activity induced by signalling pathways (Ross-Innes *et al.*, 2012). An additional possibility is alterations in the expression or function of other transcription factors such as AP1 (above) or ELF5. Based on the findings in this chapter, it is proposed that ELF5 can direct FOXA1 to new genomic binding sites, which may be a novel mechanism of altering ER binding and sensitivity to endocrine therapies in breast cancer.

### **Potential mechanisms of ELF5-driven redistribution of FOXA1 binding**

There are several possible mechanisms by which increased ELF5 expression may direct FOXA1 to new genomic regions (Figure 4.22). These were explored through a more detailed analysis of the overlap between ELF5 and FOXA1 binding, and the transcription factor motifs present in the gained and lost FOXA1 binding sites.

Firstly, ELF5 and FOXA1 may directly co-operate, with both factors binding to DNA and interacting to form a heterodimer (Figure 4.22A). Direct co-operativity between an Ets factor and a Forkhead factor has been previously demonstrated, with *Etv2* and *Foxc2* binding to a composite enhancer motif (consisting of an Ets motif and a non-canonical Forkhead motif) to regulate the expression of genes involved in endothelial development (De Val *et al.*, 2008). However, analysis of the ETS and Forkhead motifs identified in the gained FOXA1 binding sites showed that both motifs occurred within the same 100 base pair summit region in only one-third of cases. At sites where the motifs co-occurred, there was no consistent pattern to their spacing, which ranged from 0-74 base pairs. Furthermore, the ELF5 RIME experiment in Chapter 5 (investigating ELF5 protein-protein interactions) did not identify FOXA1 as an ELF5 interaction candidate. Overall, these findings suggest that direct co-operativity between ELF5 and FOXA1 as a heterodimer is unlikely.



**Figure 4.22: Potential mechanisms of ELF5-driven redistribution of FOXA1 binding**

(A) Direct co-operativity of ELF5 and FOXA1 may occur through binding to DNA regulatory regions as a heterodimer. Both factors bind to rigidly-spaced ETS and Forkhead DNA motifs or a combined motif. (B) Another mechanism of direct co-operativity is tethering; in this example, FOXA1 binds to DNA-bound ELF5. (C) Tethering may also involve both factors

binding to DNA at distal sites, which are brought into close proximity by DNA looping. Additional unknown members of a larger complex (purple) may also contribute to transcription factor tethering. (D) In assisted loading, ELF5 binding promotes an open chromatin configuration through the recruitment of chromatin remodelling factors (CR), facilitating subsequent FOXA1 binding. The dynamic nature of transcription factor binding means that simultaneous binding of both transcription factors may not be detected. In this scenario, the transcription factor motifs may be variably spaced; however, the detection of both ETS and Forkhead motifs at these sites suggests that most binding events occur within 100 base pairs. (E) ELF5 and FOXA1 may also indirectly co-operate by binding to common members of a larger complex, stabilising the binding of both factors. bp, base pairs; CR, chromatin remodeller; TSS, transcription start site.

A second possible mechanism of direct co-operativity is tethering of FOXA1 by ELF5 (Figures 4.22B and 4.22C). This is suggested by the one-third of gained FOXA1 binding sites that contain an ETS motif only and no Forkhead motif. Once again, however, the absence of FOXA1 in the ELF5 RIME argues against this scenario. A FOXA1 motif outside the 100bp summit region included in the motif analysis cannot be excluded at these sites.

Thirdly, and considered to be the most likely explanation, ELF5 may promote the binding of FOXA1 through indirect mechanisms of co-operativity. One example is the initiation of regional chromatin remodelling, a function typically associated with “pioneer factors” (Figure 4.22D). One of the classic features of pioneer factors such as FOXA1 is the intrinsic ability to open chromatin without assistance from ATP-dependent chromatin remodellers (Cirillo *et al.*, 2002), thereby restricting pioneer activity to highly specialised transcription factors. However, other studies suggest that pioneering activity might be more widespread than previously thought. A computational method to identify transcription factor binding sites from mouse embryonic stem cell (mESC) DNase I hypersensitivity data identified 120 (16% of total) transcription factor motifs that were robustly associated with an increase in chromatin accessibility from one differentiation stage to the next, consistent with pioneer activity; interestingly, these 120 pioneer motifs included 9 ETS factors (of 13 total included in the analysis) (Sherwood *et al.*, 2014). It has also been shown that ER and GR can act as pioneer factors at a subset of their binding sites (Gertz *et al.*, 2013; Voss *et al.*, 2011) and that both of these transcription factors can change the distribution of genomic FOXA1 binding by altering chromatin accessibility (Swinstead *et al.*, 2016a). Another study of unliganded ER genomic binding demonstrated that ER knockdown causes a decrease in FOXA1 binding, indicating that ER facilitates FOXA1 binding at selected sites (Caizzi *et al.*,

2014). This contrasts with previous studies in which knockdown of ER had no effect on FOXA1 binding (Hurtado *et al.*, 2011) and the reason for the discrepancy in the results of these studies is unclear. However, together these data suggest that while the primary role of FOXA1 is to enable additional factors to bind to chromatin, co-operative interactions with other factors may facilitate the binding of FOXA1 to a subset of lower affinity sites.

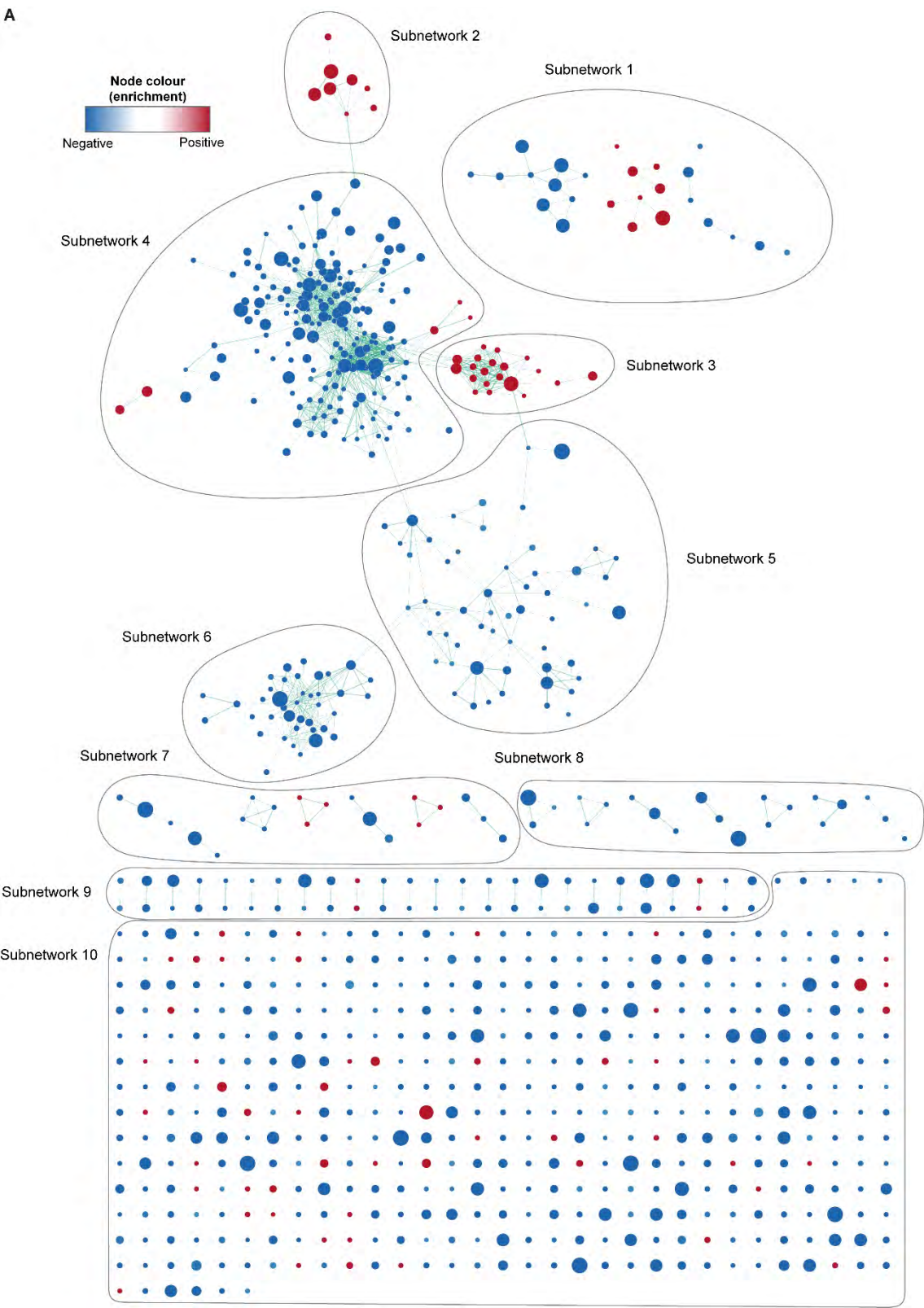
What these studies also demonstrate is that the ability of a transcription factor such as ER, GR or ELF5 to act as a pioneer factor may depend on the context of binding rather than an intrinsic property of the protein. This is expanded on in the “DynALoad” model proposed by Swinstead *et al* (2016b), in which many transcription factors may act as pioneers (or “initiators”) through the recruitment of chromatin remodelling complexes and the mechanism of dynamic assisted loading. Several pioneer factors, including GATA3 and progesterone receptor (PR), have been shown to rely on the recruitment of chromatin remodellers such as SMARCA4 (also known as BRG1, a component of the SWI/SNF complex) for their pioneer function (Ballare *et al.*, 2013; Takaku *et al.*, 2016). This model is also consistent with the recent finding that most pioneer factors are not stably recruited to chromatin but exhibit highly dynamic binding (Swinstead *et al.*, 2016a). Recruitment of common co-factors may also help to stabilise the binding of both factors (Figure 4.22E).

The mechanism of dynamic assisted loading may in part explain why there is a significant enrichment for the ELF5 motif in the gained FOXA1 binding sites and yet the level of ELF5 co-binding detected at these sites is relatively low (21%). One limitation of ChIP-seq is that it cannot capture the dynamics of transcription factor binding. In the dynamic assisted loading model, the binding of an initiating factor (in this case ELF5) creates an open chromatin conformation through the recruitment of chromatin remodellers, thereby facilitating the binding of a secondary factor (FOXA1). However, the initiating event may be too transient to be captured by ChIP-seq, a phenomenon termed “hit and run” (Voss and Hager, 2014).

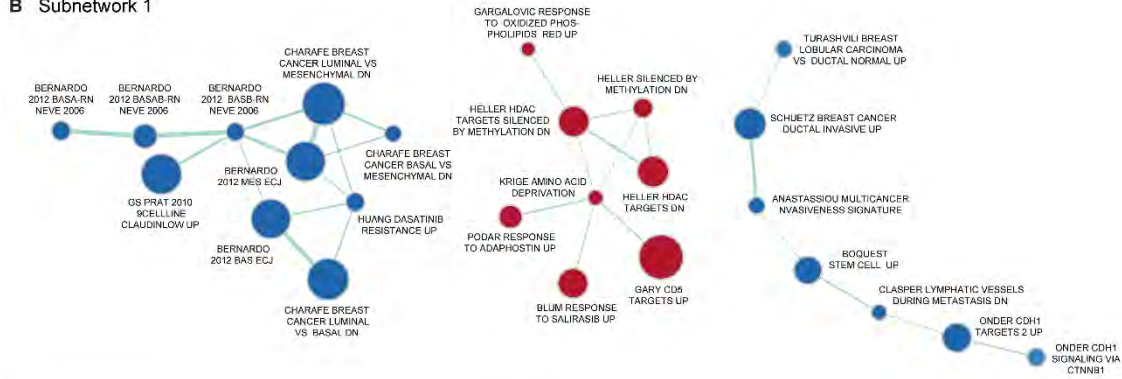


# Additional Figures and Tables

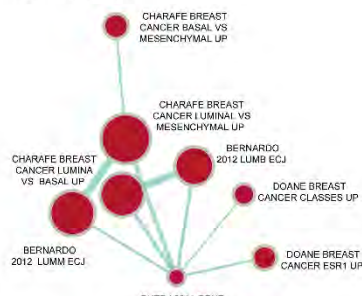
Additional Figure 4.1: Enlarged Cytoscape sub-networks (based on Figure 4.1)



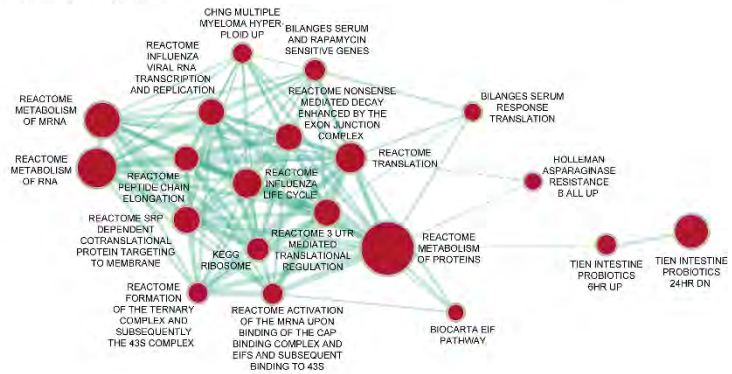
## B Subnetwork 1



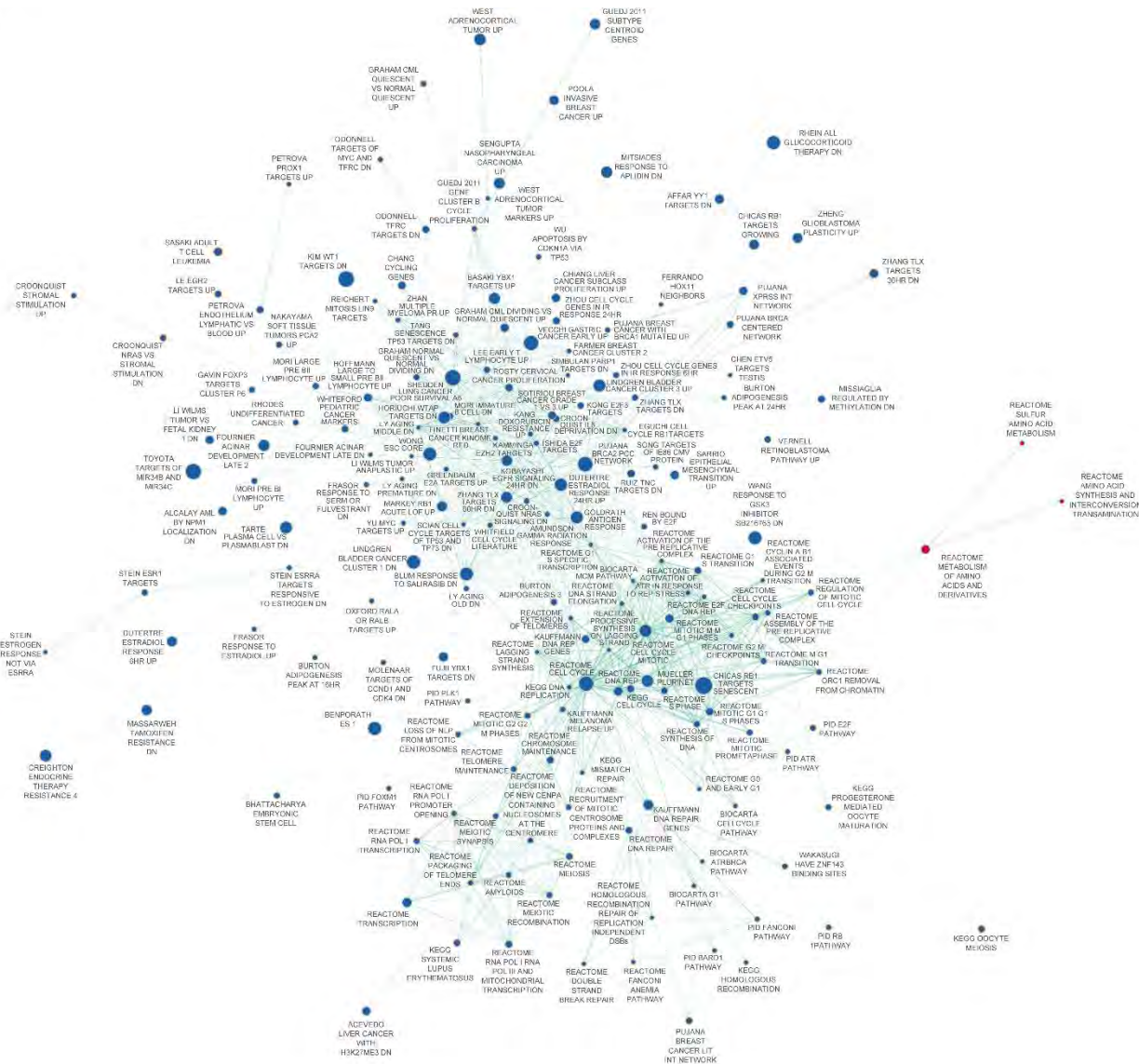
## C Subnetwork 2



## D Subnetwork 3

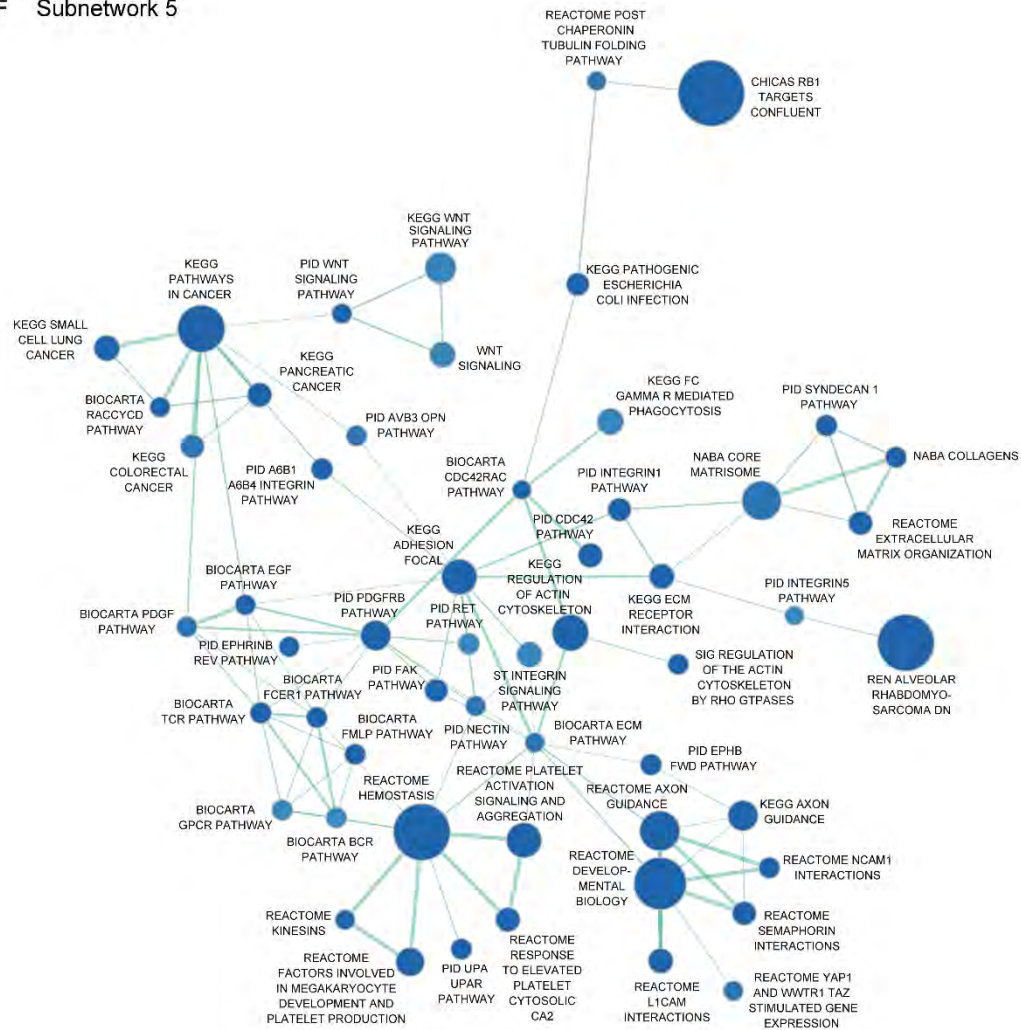


E Subnetwork 4

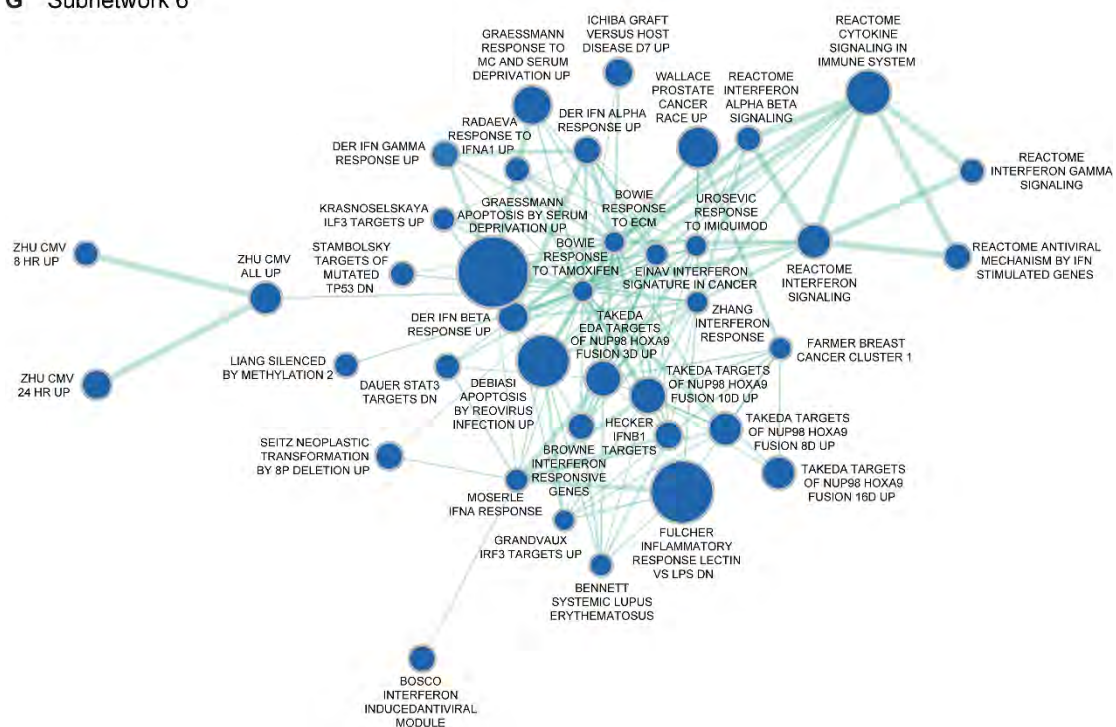




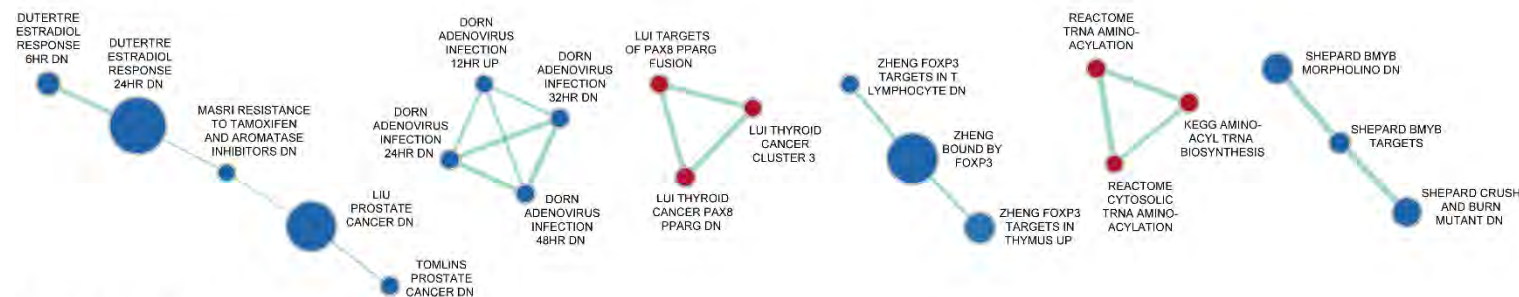
F Subnetwork 5



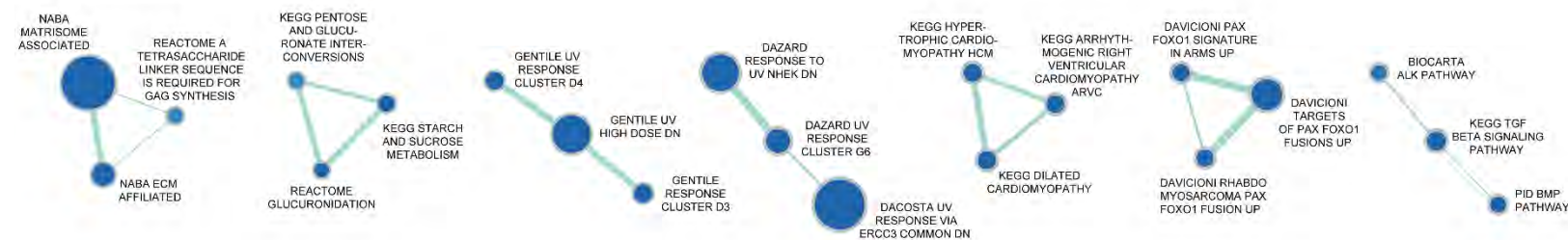
**G** Subnetwork 6



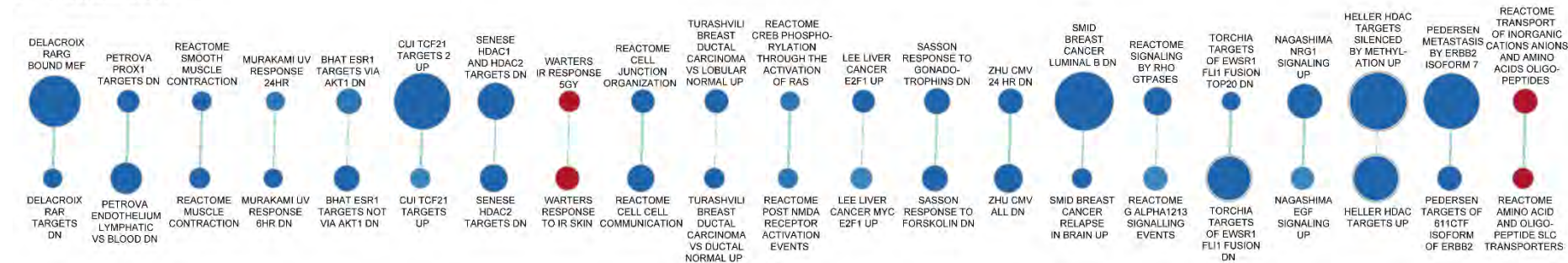
## H Subnetwork 7



## I Subnetwork 8

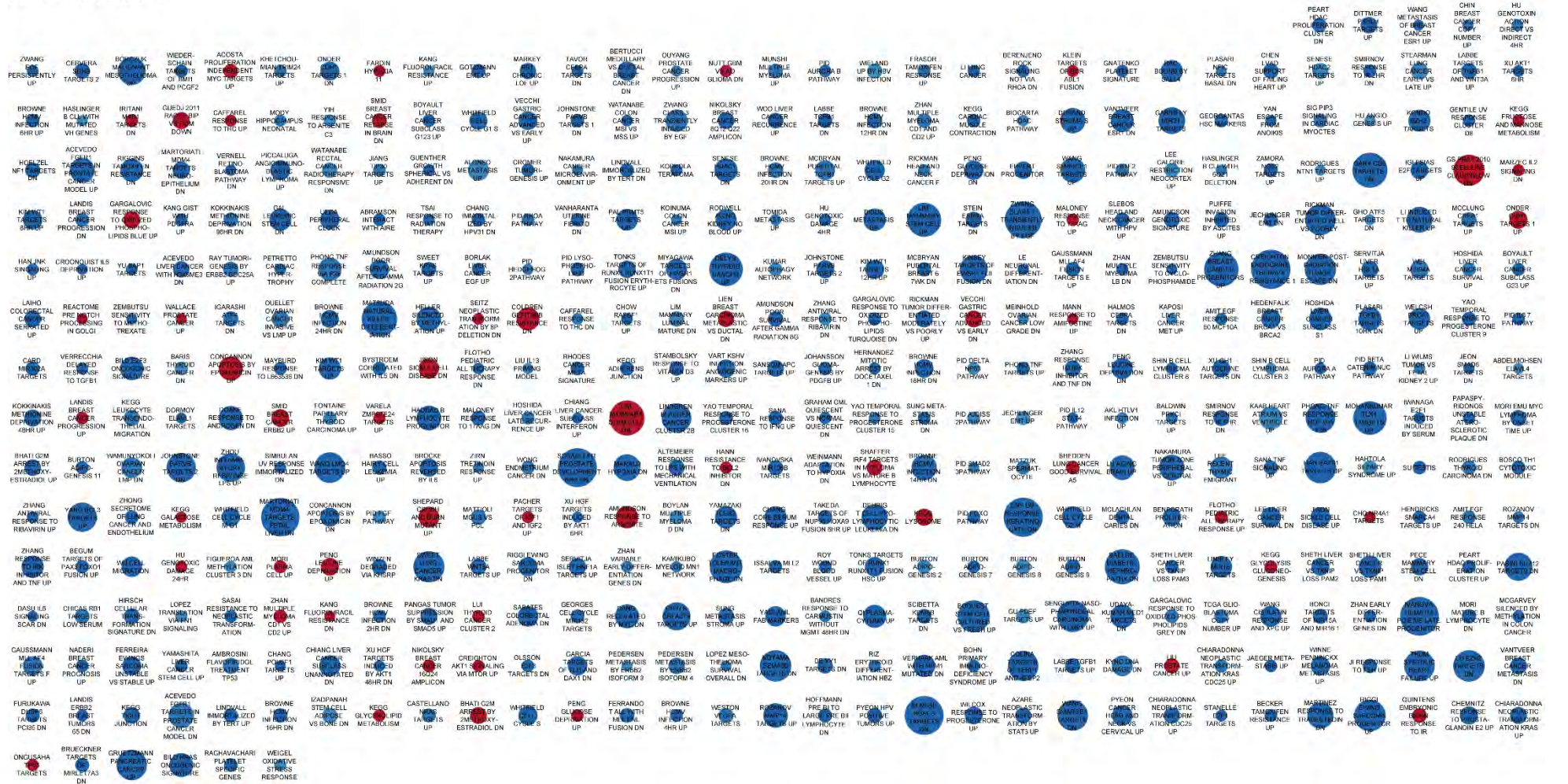


## J Subnetwork 9





## K Subnetwork 10



**Additional figure 4.1.** Sub-networks from RNA-seq GSEA Cytoscape network (Figure 4.1) comparing vehicle- and doxycycline-treated MCF7-ELF5-V5 cells. GSEA was performed using the C2 gene set collection from MSigDB with additional manually curated breast cancer sets. The circular nodes represent gene sets, with the diameter proportional to the size of the gene set. The colour of the node indicates the direction and magnitude of gene set enrichment based on normalised enrichment score following ELF5 induction (see scale), with red indicating up-regulation and blue down-regulation. Lines (“edges”) represent an overlap between connected gene sets, with the line thickness proportional to the degree of overlap. The labels summarise functional themes for prominent clusters. All gene sets with an FDR <0.05 and p-value <0.005 are shown, with the entire network containing 870 nodes and 1380 edges. An overview of the network, with sub-networks as indicated, is shown in panel (A). Sub-networks 1-10 are shown in panels (B-K).

**Additional Table 4.1: Differentially expressed genes (up-regulated) in MCF7-ELF5-V5 RNA-seq**

Gene name	Gene description	Gene type	Log2fc	FDR	ChIP
<b>ELF5</b>	E74 like ETS transcription factor 5	protein coding	3.1058	3.43E-10	
<b>COTL1</b>	coactosin like F-actin binding protein 1	protein coding	1.0013	8.23E-07	
<b>IGFBP5</b>	insulin like growth factor binding protein 5	protein coding	1.4985	1.04E-06	
<b>GLA</b>	galactosidase alpha	protein coding	1.2513	1.04E-06	
SLC9A2	solute carrier family 9 member A2	protein coding	0.8175	1.10E-06	
<b>DDIT4</b>	DNA damage inducible transcript 4	protein coding	0.8306	1.28E-06	
<b>DHRS2</b>	dehydrogenase/reductase 2	protein coding	0.9587	1.34E-06	
<b>GDF15</b>	growth differentiation factor 15	protein coding	1.4806	2.97E-06	
<b>ALDH1L2</b>	aldehyde dehydrogenase 1 family member L2	protein coding	0.8169	3.29E-06	
ARMCX1	armadillo repeat containing, X-linked 1	protein coding	1.0994	5.43E-06	
SLC4A8	solute carrier family 4 member 8	protein coding	0.7833	5.43E-06	
<b>VDR</b>	vitamin D (1,25- dihydroxyvitamin D3) receptor	protein coding	0.8182	6.33E-06	
ALDH3B2	aldehyde dehydrogenase 3 family member B2	protein coding	0.7811	8.32E-06	
<b>ATP2C2</b>	ATPase secretory pathway Ca2+ transporting 2	protein coding	0.7782	1.47E-05	
<b>MYO18A</b>	myosin XVIIIa	protein coding	1.0150	1.63E-05	
<b>SLC7A11</b>	solute carrier family 7 member 11	protein coding	0.8774	1.76E-05	
<b>PFKFB3</b>	6-phosphofructo-2-kinase/fructose-2,6-biphosphatase 3	protein coding	0.9411	1.93E-05	
TSPAN15	tetraspanin 15	protein coding	0.6761	1.97E-05	
<b>ZNF552</b>	zinc finger protein 552	protein coding	0.9529	1.99E-05	
<b>B4GALT1</b>	beta-1,4-galactosyltransferase 1	protein coding	0.9615	2.13E-05	
<b>DEPTOR</b>	DEP domain containing MTOR interacting protein	protein coding	1.2595	2.16E-05	
<b>ATP2A3</b>	ATPase sarcoplasmic/endoplasmic reticulum Ca2+ transporting 3	protein coding	1.3306	2.38E-05	
PXMP4	peroxisomal membrane protein 4	protein coding	0.7528	2.39E-05	
CST4	cystatin S	protein coding	5.2474	2.85E-05	
PXDN	peroxidase	protein coding	0.7777	2.85E-05	
PLBD1	phospholipase B domain containing 1	protein coding	0.7088	2.88E-05	
<b>PLEKHS1</b>	pleckstrin homology domain containing S1	protein coding	1.3174	2.96E-05	
<b>GRAMD2</b>	GRAM domain containing 2	protein coding	2.5339	3.32E-05	
GRPR	gastrin releasing peptide receptor	protein coding	1.3032	3.92E-05	
<b>NUPR1</b>	nuclear protein 1, transcriptional regulator	protein coding	1.1589	4.37E-05	
CUEDC1	CUE domain containing 1	protein coding	0.6189	4.40E-05	
<b>ASAH1</b>	N-acylsphingosine amidohydrolase 1	protein coding	0.6289	4.42E-05	
<b>EIF4EBP1</b>	eukaryotic translation initiation factor 4E binding protein 1	protein coding	0.6551	4.59E-05	
TMC5	transmembrane channel like 5	protein coding	1.4069	4.81E-05	
<b>LAMP3</b>	lysosomal associated membrane protein 3	protein coding	1.9013	5.14E-05	
<b>RPS6KA5</b>	ribosomal protein S6 kinase A5	protein coding	0.8222	5.14E-05	
STARD13	StAR related lipid transfer domain containing 13	protein coding	0.5996	6.27E-05	
<b>BCAS1</b>	breast carcinoma amplified sequence 1	protein coding	1.3854	6.82E-05	
<b>VTGN1</b>	V-set domain containing T-cell activation inhibitor 1	protein coding	3.8416	7.15E-05	
<b>FAM107B</b>	family with sequence similarity 107 member B	protein coding	1.1434	8.47E-05	
EHF	ETS homologous factor	protein coding	0.9806	9.75E-05	
<b>ASNS</b>	asparagine synthetase (glutamine-hydrolyzing)	protein coding	0.5871	9.97E-05	



<b>ZFAS1</b>	ZNF1 antisense RNA 1	antisense	0.6949	1.05E-04	
DPH1	diphthamide biosynthesis 1	protein coding	0.7930	1.18E-04	
S100A9	S100 calcium binding protein A9	protein coding	1.6366	1.22E-04	
RAB17	RAB17, member RAS oncogene family	protein coding	0.7064	1.47E-04	
C6orf141	chromosome 6 open reading frame 141	protein coding	0.7378	1.93E-04	
CORO2A	coronin 2A	protein coding	1.1699	1.97E-04	
ITPR1	inositol 1,4,5-trisphosphate receptor type 1	protein coding	0.6051	2.03E-04	
CYP4Z2P	cytochrome P450 family 4 subfamily Z member 2, pseudogene	pseudogene (TU)	3.4599	2.16E-04	
PHLDA1	pleckstrin homology like domain family A member 1	protein coding	0.5874	2.21E-04	
<b>CCDC88C</b>	coiled-coil domain containing 88C	protein coding	0.7075	2.25E-04	
CYP4Z1	cytochrome P450 family 4 subfamily Z member 1	protein coding	5.7088	2.30E-04	
IVL	involucrin	protein coding	4.0436	2.38E-04	
PLA2G16	phospholipase A2 group XVI	protein coding	0.9932	2.60E-04	
SH2D3C	SH2 domain containing 3C	protein coding	4.2446	2.76E-04	
SRGAP3	SLIT-ROBO Rho GTPase activating protein 3	protein coding	0.8034	2.95E-04	
GALM	galactose mutarotase	protein coding	1.7545	3.03E-04	
C14orf159	chromosome 14 open reading frame 159	protein coding	1.9688	3.12E-04	
PGPEP1	pyroglutamyl-peptidase I	protein coding	0.6047	3.14E-04	
<b>SAMD12</b>	sterile alpha motif domain containing 12	protein coding	1.8437	3.57E-04	
<b>MFS2A</b>	major facilitator superfamily domain containing 2A	protein coding	1.1812	3.80E-04	
FBXO27	F-box protein 27	protein coding	0.5976	3.99E-04	
OTUB2	OTU deubiquitinase, ubiquitin aldehyde binding 2	protein coding	1.1267	4.12E-04	
RNF144B	ring finger protein 144B	protein coding	0.7791	4.21E-04	
<b>ARHGDI1</b>	Rho GDP dissociation inhibitor beta	protein coding	1.8888	4.33E-04	
<b>FAS</b>	Fas cell surface death receptor	protein coding	0.8548	5.06E-04	
<b>SDCBP</b>	syndecan binding protein	protein coding	0.7227	5.10E-04	
LINC01133	long intergenic non-protein coding RNA 1133	lincRNA	4.0034	5.24E-04	
AZGP1	alpha-2-glycoprotein 1, zinc-binding	protein coding	3.1338	5.94E-04	
NABP1	nucleic acid binding protein 1	protein coding	0.7226	6.05E-04	
KCNK6	potassium two pore domain channel subfamily K member 6	protein coding	0.6436	6.17E-04	
MYB	MYB proto-oncogene, transcription factor	protein coding	0.6876	6.29E-04	
SLFN5	schlafen family member 5	protein coding	1.7863	6.78E-04	
DEGS2	delta 4-desaturase, sphingolipid 2	protein coding	1.1826	7.35E-04	
PRRG1	proline rich and Gla domain 1	protein coding	1.8772	8.43E-04	
CD22	CD22 molecule	protein coding	1.8141	8.58E-04	
FHOD1	formin homology 2 domain containing 1	protein coding	0.5941	8.72E-04	
CHST1	carbohydrate sulfotransferase 1	protein coding	3.1761	9.12E-04	
<b>CRISP3</b>	cysteine rich secretory protein 3	protein coding	1.2172	9.76E-04	
<b>CD82</b>	CD82 molecule	protein coding	0.9500	1.00E-03	
VWF	von Willebrand factor	protein coding	2.3465	1.09E-03	
<i>AL365205.1</i>	ENSG00000124593 Prickle-like protein 4	protein coding	0.8003	1.17E-03	
AQP3	aquaporin 3 (Gill blood group)	protein coding	2.0948	1.19E-03	
MME	membrane metalloendopeptidase	protein coding	1.3198	1.19E-03	
SMPDL3B	sphingomyelin phosphodiesterase acid like 3B	protein coding	0.8784	1.27E-03	
<b>GRHL3</b>	grainyhead like transcription factor 3	protein coding	1.1962	1.29E-03	
ELL3	elongation factor for RNA polymerase II 3	protein coding	0.6749	1.36E-03	

TMEM92	transmembrane protein 92	protein coding	3.4122	1.36E-03	
<b>NTN4</b>	netrin 4	protein coding	1.0964	1.59E-03	
FUT3	fucosyltransferase 3 (Lewis blood group)	protein coding	4.0669	1.64E-03	
FBLN5	fibulin 5	protein coding	3.3166	1.65E-03	
ARC	activity regulated cytoskeleton associated protein	protein coding	1.8474	1.69E-03	
C2CD2L	C2CD2 like	protein coding	0.6356	1.71E-03	
CAMP	cathelicidin antimicrobial peptide	protein coding	3.3731	1.89E-03	
CPAMD8	C3 and PZP like, alpha-2-macroglobulin domain containing 8	protein coding	0.7200	2.07E-03	
<b>BLNK</b>	B-cell linker	protein coding	2.2669	2.20E-03	
<b>YPEL2</b>	yippee like 2	protein coding	0.7395	2.24E-03	
ZP3	zona pellucida glycoprotein 3	protein coding	0.7391	2.41E-03	
ARMCX2	armadillo repeat containing, X-linked 2	protein coding	1.0045	2.46E-03	
GGT1	gamma-glutamyltransferase 1	protein coding	1.3980	2.65E-03	
SOWAHB	sosondowah ankyrin repeat domain family member B	protein coding	0.7709	2.72E-03	
<b>LCN2</b>	lipocalin 2	protein coding	1.2119	2.76E-03	
<b>RND1</b>	Rho family GTPase 1	protein coding	1.3252	2.98E-03	
ITGAL	integrin subunit alpha L	protein coding	3.5359	3.09E-03	
<b>KMO</b>	kynurenine 3-monooxygenase	protein coding	3.2994	3.20E-03	
RHOV	ras homolog family member V	protein coding	0.5907	3.20E-03	
<b>GRIK3</b>	glutamate ionotropic receptor kainate type subunit 3	protein coding	0.7069	3.24E-03	
BBC3	BCL2 binding component 3	protein coding	0.8029	3.58E-03	
NDRG4	NDRG family member 4	protein coding	0.7257	3.58E-03	
LRRC4	leucine rich repeat containing 4	protein coding	1.7952	3.74E-03	
<b>TFF1</b>	trefoil factor 1	protein coding	0.6745	3.74E-03	
PADI3	peptidyl arginine deiminase 3	protein coding	2.3825	4.33E-03	
PIP	prolactin induced protein	protein coding	3.2859	4.36E-03	
IQCH-AS1	IQCH antisense RNA 1	lincRNA	0.6805	4.60E-03	
<b>CYFIP2</b>	cytoplasmic FMR1 interacting protein 2	protein coding	0.6974	4.74E-03	
ITGB2	integrin subunit beta 2	protein coding	2.4397	4.76E-03	
ST6GALNAC4	ST6 N-acetylgalactosaminide alpha-2,6-sialyltransferase 4	protein coding	0.7072	5.25E-03	
UCA1	urothelial cancer associated 1 (non-protein coding)	processed transcript	3.0133	5.29E-03	
CLSTN3	calsyntenin 3	protein coding	0.8906	5.38E-03	
<b>MGLL</b>	monoglyceride lipase	protein coding	0.9708	5.53E-03	
<b>ULBP1</b>	UL16 binding protein 1	protein coding	1.1207	5.72E-03	
AL390198.2	ENSG00000250090	pseudogene (U)	2.9603	5.93E-03	
ARMCX4	armadillo repeat containing, X-linked 4	protein coding	0.8577	6.19E-03	
CALCOCO1	calcium binding and coiled-coil domain 1	protein coding	0.6286	6.21E-03	
AL121839.2	ENSG00000260711	sense intronic	0.6835	6.24E-03	
TMEM150A	transmembrane protein 150A	protein coding	0.6420	6.31E-03	
<b>MGP</b>	matrix Gla protein	protein coding	0.6221	6.35E-03	
CAMKK1	calcium/calmodulin dependent protein kinase kinase 1	protein coding	0.9133	6.69E-03	
SPP1	secreted phosphoprotein 1	protein coding	3.0130	6.72E-03	
NDUFA4L2	NDUFA4, mitochondrial complex associated like 2	protein coding	1.3074	6.98E-03	
ADM2	adrenomedullin 2	protein coding	0.6803	7.25E-03	
ACTL10	actin like 10	protein coding	0.8670	7.29E-03	

AC008763.2	ENSG00000268400	protein coding	3.4223	7.43E-03	
PRX	periaxin	protein coding	0.7361	7.43E-03	
CAPN8	calpain 8	protein coding	2.3682	7.58E-03	
DLX1	distal-less homeobox 1	protein coding	0.6852	7.68E-03	
SCNN1A	sodium channel epithelial 1 alpha subunit	protein coding	0.6297	7.90E-03	
ALDH1A3	aldehyde dehydrogenase 1 family member A3	protein coding	0.6680	7.96E-03	
PTGS1	prostaglandin-endoperoxide synthase 1	protein coding	1.9297	8.26E-03	
<b>S100P</b>	S100 calcium binding protein P	protein coding	1.0890	8.81E-03	
<b>PMAIP1</b>	phorbol-12-myristate-13-acetate-induced protein 1	protein coding	0.7347	9.32E-03	
GJC3	gap junction protein gamma 3	protein coding	2.7993	9.58E-03	
PTK2B	protein tyrosine kinase 2 beta	protein coding	0.7125	9.69E-03	
TTBK1	tau tubulin kinase 1	protein coding	2.3081	0.0100	
NDRG2	NDRG family member 2	protein coding	0.8343	0.0100	
CDH5	cadherin 5	protein coding	2.9531	0.0104	
<b>MACC1</b>	MACC1, MET transcriptional regulator	protein coding	1.2743	0.0115	
<b>HMOX1</b>	heme oxygenase 1	protein coding	0.7509	0.0115	
<b>GLRX</b>	glutaredoxin	protein coding	1.1755	0.0117	
SPNS2	sphingolipid transporter 2	protein coding	1.4531	0.0117	
AC018755.3	ENSG00000269388	pseudogene (TP)	2.7208	0.0117	
GTF2IRD2P1	GTF2I repeat domain containing 2 pseudogene 1	pseudogene (TU)	2.6326	0.0119	
ABCA4	ATP binding cassette subfamily A member 4	protein coding	1.9647	0.0120	
FAM184A	family with sequence similarity 184 member A	protein coding	0.7213	0.0120	
TNS2	tensin 2	protein coding	0.6169	0.0121	
CD36	CD36 molecule	protein coding	1.0323	0.0126	
AL021546.1	ENSG00000111780	protein coding	3.3340	0.0133	
GPSM3	G protein signaling modulator 3	protein coding	2.3187	0.0137	
AC138811.2	ENSG00000260342	protein coding	3.5472	0.0141	
CDC42EP5	CDC42 effector protein 5	protein coding	1.0405	0.0142	
RASA4	RAS p21 protein activator 4	protein coding	0.6887	0.0142	
SNRPN	small nuclear ribonucleoprotein polypeptide N	protein coding	1.8808	0.0151	
EPHA10	EPH receptor A10	protein coding	0.6110	0.0161	
HRK	harakiri, BCL2 interacting protein	protein coding	0.7819	0.0168	
KLF15	Kruppel like factor 15	protein coding	1.6416	0.0168	
SYBU	syntabulin	protein coding	0.5914	0.0180	
AC023157.3	ENSG00000276900	antisense	0.7969	0.0185	
<b>VGLL1</b>	vestigial like family member 1	protein coding	1.6547	0.0187	
AL365181.3	ENSG00000272405	antisense	1.6275	0.0188	
PRSS1	protease, serine 1	protein coding	2.5377	0.0188	
MRV11	murine retrovirus integration site 1 homolog	protein coding	2.7587	0.0190	
LINC00173	long intergenic non-protein coding RNA 173	processed transcript	1.3886	0.0195	
ZNF284	zinc finger protein 284	protein coding	0.9578	0.0195	
ENG	endoglin	protein coding	2.4571	0.0196	
CASP9	caspase 9	protein coding	0.6386	0.0199	
LMO2	LIM domain only 2	protein coding	2.0045	0.0204	
RGPD4	RANBP2-like and GRIP domain containing 4	protein coding	2.5896	0.0207	
B3GNT3	UDP-GlcNAc:betaGal beta-1,3-N-acetylglucosaminyltransferase 3	protein coding	1.0894	0.0208	

FAM86JP	family with sequence similarity 86 member J, pseudogene	pseudogene (TU)	0.6672	0.0211	
BTN3A1	butyrophilin subfamily 3 member A1	protein coding	0.9569	0.0213	
BPGM	bisphosphoglycerate mutase	protein coding	0.5907	0.0213	
LRG1	leucine rich alpha-2-glycoprotein 1	protein coding	1.6723	0.0213	
DISP3	dispatched RND transporter family member 3	protein coding	1.2251	0.0219	
SPATA18	spermatogenesis associated 18	protein coding	0.7773	0.0225	
AL583839.1	ENSG00000227603	antisense	2.5866	0.0227	
NEIL1	nei like DNA glycosylase 1	protein coding	0.6011	0.0234	
FCGBP	Fc fragment of IgG binding protein	protein coding	2.1568	0.0236	
TMEM40	transmembrane protein 40	protein coding	1.7675	0.0237	
MSMB	microseminoprotein beta	protein coding	1.8795	0.0248	
ACTL8	actin like 8	protein coding	0.8535	0.0250	
AKR1C2	aldo-keto reductase family 1 member C2	protein coding	0.6691	0.0253	
PFKFB4	6-phosphofructo-2-kinase/fructose-2,6-biphosphatase 4	protein coding	0.6128	0.0256	
CRISP1	cysteine rich secretory protein 1	protein coding	1.8737	0.0256	
PDE2A	phosphodiesterase 2A	protein coding	2.7420	0.0261	
RFTN1	raftlin, lipid raft linker 1	protein coding	0.8308	0.0268	
PABPC1L2B-AS1	PABPC1L2B antisense RNA 1 (head to head)	antisense	3.3787	0.0269	
FLT4	fms related tyrosine kinase 4	protein coding	0.6172	0.0270	
HEATR4	HEAT repeat containing 4	protein coding	2.4940	0.0272	
AC079601.1	ENSG00000257225	antisense	2.2942	0.0277	
SMIM6	small integral membrane protein 6	protein coding	2.2942	0.0277	
AC069368.1	ENSG00000249240	protein coding	2.7488	0.0281	
AL390195.1	ENSG00000243960	sense overlapping	3.9554	0.0282	
ASPHD1	aspartate beta-hydroxylase domain containing 1	protein coding	0.8153	0.0284	
COLEC11	collectin subfamily member 11	protein coding	2.4018	0.0284	
ARHGEF26-AS1	ARHGEF26 antisense RNA 1	processed transcript	0.8559	0.0297	
<b>CAPG</b>	capping actin protein, gelsolin like	protein coding	0.8685	0.0299	
AC021739.5	ENSG00000259762	antisense	2.4080	0.0300	
FA2H	fatty acid 2-hydroxylase	protein coding	0.7457	0.0302	
<b>CLGN</b>	calmegin	protein coding	1.2288	0.0311	
SH3PXD2A	SH3 and PX domains 2A	protein coding	1.0300	0.0313	
<b>CHAC1</b>	ChaC glutathione specific gamma-glutamylcyclotransferase 1	protein coding	0.5875	0.0327	
VMO1	vitelline membrane outer layer 1 homolog	protein coding	1.2668	0.0329	
STAMBPL1	STAM binding protein like 1	protein coding	0.6016	0.0331	
DNASE2B	deoxyribonuclease 2 beta	protein coding	2.5627	0.0331	
ADGRE5	adhesion G protein-coupled receptor E5	protein coding	0.6176	0.0332	
SELENOP	selenoprotein P	protein coding	0.9014	0.0343	
TIAF1	TGFB1-induced anti-apoptotic factor 1	protein coding	0.8721	0.0358	
KCNE4	potassium voltage-gated channel subfamily E regulatory subunit 4	protein coding	1.6596	0.0363	
RPL21P119	ribosomal protein L21 pseudogene 119	pseudogene (P)	1.6122	0.0363	
AL136531.2	ENSG00000274322	protein coding	2.4936	0.0366	
AC091230.1	ENSG00000261606	processed transcript	2.6671	0.0370	

<i>AC009102.2</i>	ENSG00000272372	antisense	2.2161	0.0373	
PIWIL2	piwi like RNA-mediated gene silencing 2	protein coding	2.2762	0.0378	
LRRC37A4P	leucine rich repeat containing 37 member A4, pseudogene	pseudogene (TU)	0.6101	0.0378	
NWD1	NACHT and WD repeat domain containing 1	protein coding	1.1046	0.0379	
MAGEA2B	MAGE family member A2B	protein coding	3.0574	0.0381	
CCDC160	coiled-coil domain containing 160	protein coding	2.0061	0.0383	
LTBP2	latent transforming growth factor beta binding protein 2	protein coding	1.1795	0.0383	
MAPT	microtubule associated protein tau	protein coding	0.6244	0.0390	
SNORA63	ENSG00000200418 Small nucleolar RNA SNORA63	snoRNA	2.3722	0.0399	
PAX5	paired box 5	protein coding	0.6165	0.0415	
RPL32P29	ribosomal protein L32 pseudogene 29	pseudogene (P)	0.7946	0.0418	
LINC01703	long intergenic non-protein coding RNA 1703	lincRNA	1.3053	0.0419	
SLC13A3	solute carrier family 13 member 3	protein coding	1.0061	0.0424	
LINC02361*	long intergenic non-protein coding RNA 2361	lincRNA	1.8111	0.0425	
<i>AC005091.1</i>	ENSG00000229893	antisense	2.1215	0.0428	
<i>AL359643.2</i>	ENSG00000271978	lincRNA	2.1215	0.0428	
VAMP5	vesicle associated membrane protein 5	protein coding	1.3955	0.0446	
<i>AC010834.2</i>	ENSG00000253848	antisense	2.2802	0.0447	
NME1-NME2	NME1-NME2 readthrough	protein coding	0.6364	0.0449	
<i>AL359232.1</i>	ENSG00000258561	lincRNA	2.2368	0.0450	
LDLRAD2	low density lipoprotein receptor class A domain containing 2	protein coding	2.1770	0.0450	
C1orf168	chromosome 1 open reading frame 168	protein coding	2.2880	0.0451	
SRGAP2D	SLIT-ROBO Rho GTPase activating protein 2D (pseudogene)	pseudogene (U)	2.9961	0.0455	
ACVRL1	activin A receptor like type 1	protein coding	2.4499	0.0465	
GAL3ST1	galactose-3-O-sulfotransferase 1	protein coding	2.4788	0.0465	
POLR2J2	RNA polymerase II subunit J2	protein coding	1.8641	0.0466	
SIRPA	signal regulatory protein alpha	protein coding	0.5910	0.0477	
PDZK1	PDZ domain containing 1	protein coding	0.6532	0.0487	
PLCXD2	phosphatidylinositol specific phospholipase C X domain containing 2	protein coding	2.1817	0.0488	
<i>AP005717.1</i>	ENSG00000245330	lincRNA	2.2739	0.0489	
GGT5	gamma-glutamyltransferase 5	protein coding	2.1749	0.0492	
LINC01088	long intergenic non-protein coding RNA 1088	antisense	2.1749	0.0492	
<i>AL355916.1</i>	ENSG00000250548	lincRNA	2.1749	0.0492	
<i>AC025283.2</i>	ENSG00000262621	protein coding	1.2299	0.0493	

**Additional Table 4.1.** List of 256 genes up-regulated following ELF5 induction, with absolute fold change >1.5 and FDR <0.05, sorted by FDR. “ChIP” column indicates the presence (purple) or absence (white) of an ELF5 ChIP-seq peak within 10kb of the TSS. Gene names in red typeface (59) were also found to be up-regulated in the MCF7-ELF5-V5 microarray experiment. Gene names in italics indicate they have no official HGNC gene symbol; Ensembl gene names and identifiers (beginning ‘ENSG’) are provided for these genes. One gene (LINC02361) is marked with an asterisk (\*), indicating that it did not have an HGNC symbol when the data were originally analysed but has since been updated. P, processed; TP, transcribed processed; TU, transcribed unprocessed; U, unprocessed.

**Additional Table 4.2: Differentially expressed genes (down-regulated) in MCF7-ELF5-V5 RNA-seq**

Gene name	Gene description	Gene type	Log2fc	FDR	ChIP
<b>DKK1</b>	dickkopf WNT signaling pathway inhibitor 1	protein coding	-1.3895	7.13E-07	
<b>UGDH</b>	UDP-glucose 6-dehydrogenase	protein coding	-0.7552	7.54E-07	
<b>DDX60</b>	DEAD/H-box helicase 60	protein coding	-1.7764	8.23E-07	
<b>ST8SIA4</b>	ST8 alpha-N-acetyl-neuraminide alpha-2,8-sialyltransferase 4	protein coding	-0.9699	1.04E-06	
<b>CLDN1</b>	claudin 1	protein coding	-1.1848	1.04E-06	
<b>COL3A1</b>	collagen type III alpha 1 chain	protein coding	-0.9028	1.04E-06	
<b>GPC6</b>	glypican 6	protein coding	-0.9340	1.04E-06	
<b>IFIT1</b>	interferon induced protein with tetratricopeptide repeats 1	protein coding	-2.6276	1.04E-06	
ISG15	ISG15 ubiquitin-like modifier	protein coding	-1.5818	1.04E-06	
IRF9	interferon regulatory factor 9	protein coding	-1.6535	1.10E-06	
<b>RDX</b>	radixin	protein coding	-0.8153	1.48E-06	
<b>NRP1</b>	neuropilin 1	protein coding	-0.7963	2.03E-06	
OASL	2'-5'-oligoadenylate synthetase like	protein coding	-1.5814	2.07E-06	
<b>SNAI2</b>	snail family transcriptional repressor 2	protein coding	-2.1851	3.13E-06	
OAS2	2'-5'-oligoadenylate synthetase 2	protein coding	-2.4286	3.59E-06	
FAM84A	family with sequence similarity 84 member A	protein coding	-0.9015	3.59E-06	
CWC27	CWC27 spliceosome associated protein homolog	protein coding	-0.8849	5.27E-06	
<b>ST3GAL1</b>	ST3 beta-galactoside alpha-2,3-sialyltransferase 1	protein coding	-0.6950	5.43E-06	
<b>EFEMP1</b>	EGF containing fibulin like extracellular matrix protein 1	protein coding	-0.6964	5.43E-06	
<b>IFIT2</b>	interferon induced protein with tetratricopeptide repeats 2	protein coding	-1.4504	5.43E-06	
IFI6	interferon alpha inducible protein 6	protein coding	-1.7008	6.61E-06	
<b>DDX58</b>	DEAD/H-box helicase 58	protein coding	-1.3537	7.47E-06	
<b>SLITRK6</b>	SLIT and NTRK like family member 6	protein coding	-0.8738	7.64E-06	
<b>KLHL5</b>	kelch like family member 5	protein coding	-0.7059	8.48E-06	
<b>NECAB1</b>	N-terminal EF-hand calcium binding protein 1	protein coding	-1.0843	8.77E-06	
<b>LINC00052</b>	long intergenic non-protein coding RNA 52	lincRNA	-1.9771	1.11E-05	
<b>COL12A1</b>	collagen type XII alpha 1 chain	protein coding	-0.8603	1.28E-05	
<b>AGR2</b>	anterior gradient 2, protein disulphide isomerase family member	protein coding	-0.7435	1.33E-05	
OAS3	2'-5'-oligoadenylate synthetase 3	protein coding	-0.7591	1.47E-05	
<b>EHHADH</b>	enoyl-CoA hydratase and 3-hydroxyacyl CoA dehydrogenase	protein coding	-0.9231	1.48E-05	
<b>NPNT</b>	nephronectin	protein coding	-0.7558	1.52E-05	
<b>TRPC6</b>	transient receptor potential cation channel subfamily C member 6	protein coding	-1.5780	1.63E-05	
TMEM44	transmembrane protein 44	protein coding	-1.4774	1.80E-05	
XAF1	XIAP associated factor 1	protein coding	-2.6876	1.93E-05	
GSTM3	glutathione S-transferase mu 3	protein coding	-0.6806	2.39E-05	
<b>GSTM4</b>	glutathione S-transferase mu 4	protein coding	-1.0431	2.54E-05	
<b>CACNG4</b>	calcium voltage-gated channel auxiliary subunit gamma 4	protein coding	-0.8965	2.91E-05	
ETV5	ETS variant 5	protein coding	-1.0779	2.96E-05	
IFIT3	interferon induced protein with tetratricopeptide repeats 3	protein coding	-1.6230	3.16E-05	
<b>PRICKLE1</b>	prickle planar cell polarity protein 1	protein coding	-1.0791	3.16E-05	
<b>PGM2L1</b>	phosphoglucomutase 2 like 1	protein coding	-1.2602	3.70E-05	

<b>SCIN</b>	scinderin	protein coding	-0.6818	3.71E-05	
<b>MID1</b>	midline 1	protein coding	-0.9494	3.92E-05	
<b>TFPI</b>	tissue factor pathway inhibitor	protein coding	-1.0159	4.11E-05	
<b>BMP7</b>	bone morphogenetic protein 7	protein coding	-0.6770	4.40E-05	
<b>HERC5</b>	HECT and RLD domain containing E3 ubiquitin protein ligase 5	protein coding	-1.1799	4.40E-05	
<b>DOCK10</b>	dedicator of cytokinesis 10	protein coding	-0.6278	4.81E-05	
<b>KANK2</b>	KN motif and ankyrin repeat domains 2	protein coding	-0.7261	4.81E-05	
<b>PDE5A</b>	phosphodiesterase 5A	protein coding	-0.8269	4.93E-05	
<b>SEMA3D</b>	semaphorin 3D	protein coding	-1.0863	4.95E-05	
<b>COL5A1</b>	collagen type V alpha 1 chain	protein coding	-0.7506	4.96E-05	
<b>RAB27B</b>	RAB27B, member RAS oncogene family	protein coding	-0.6026	5.02E-05	
<b>MALRD1</b>	MAM and LDL receptor class A domain containing 1	protein coding	-0.8150	5.14E-05	
<b>ASB9</b>	ankyrin repeat and SOCS box containing 9	protein coding	-1.2674	5.17E-05	
<b>ST8SIA1</b>	ST8 alpha-N-acetyl-neuraminide alpha-2,8-sialyltransferase 1	protein coding	-2.0978	5.71E-05	
<b>FILIP1L</b>	filamin A interacting protein 1 like	protein coding	-2.2029	6.86E-05	
<b>RASD1</b>	ras related dexamethasone induced 1	protein coding	-0.7178	7.70E-05	
<b>LOXL2</b>	lysyl oxidase like 2	protein coding	-1.5969	8.83E-05	
<b>LACTB</b>	lactamase beta	protein coding	-0.6367	9.97E-05	
<b>SAMD9</b>	sterile alpha motif domain containing 9	protein coding	-2.7303	9.97E-05	
<b>SLC18B1</b>	solute carrier family 18 member B1	protein coding	-0.8310	9.99E-05	
<b>DMRTA1</b>	DMRT like family A1	protein coding	-0.7772	9.99E-05	
<b>ATP2B4</b>	ATPase plasma membrane Ca2+ transporting 4	protein coding	-0.8877	1.01E-04	
<b>PLSCR1</b>	phospholipid scramblase 1	protein coding	-0.7266	1.03E-04	
<b>PRRT3</b>	proline rich transmembrane protein 3	protein coding	-0.9363	1.16E-04	
<b>RPS6KA3</b>	ribosomal protein S6 kinase A3	protein coding	-0.6769	1.28E-04	
<b>HEG1</b>	heart development protein with EGF like domains 1	protein coding	-0.9736	1.52E-04	
<b>KCNH7</b>	potassium voltage-gated channel subfamily H member 7	protein coding	-0.9759	1.52E-04	
<b>CPM</b>	carboxypeptidase M	protein coding	-1.7985	1.64E-04	
<b>KRT80</b>	keratin 80	protein coding	-0.5982	1.87E-04	
<b>KAT2B</b>	lysine acetyltransferase 2B	protein coding	-0.7178	1.87E-04	
<b>CFL2</b>	cofilin 2	protein coding	-0.8000	1.87E-04	
<b>LYPD6B</b>	LY6/PLAUR domain containing 6B	protein coding	-0.7188	1.94E-04	
<b>AKAP5</b>	A-kinase anchoring protein 5	protein coding	-0.7895	1.97E-04	
<b>PAPSS2</b>	3'-phosphoadenosine 5'-phosphosulfate synthase 2	protein coding	-0.5990	2.25E-04	
<b>RHOBTB2</b>	Rho related BTB domain containing 2	protein coding	-0.6284	2.42E-04	
<b>AC103770.1</b>	ENSG00000254251	antisense	-1.1195	2.49E-04	
<b>CAMK2N1</b>	calcium/calmodulin dependent protein kinase II inhibitor 1	protein coding	-0.6133	2.53E-04	
<b>ANKRD1</b>	ankyrin repeat domain 1	protein coding	-1.1073	2.55E-04	
<b>NRCAM</b>	neuronal cell adhesion molecule	protein coding	-0.6433	2.59E-04	
<b>HERC6</b>	HECT and RLD domain containing E3 ubiquitin protein ligase family member 6	protein coding	-0.6053	2.60E-04	
<b>SKIDA1</b>	SKI/DACH domain containing 1	protein coding	-0.8085	3.20E-04	
<b>CCL5</b>	C-C motif chemokine ligand 5	protein coding	-1.8865	3.26E-04	
<b>MMP16</b>	matrix metalloproteinase 16	protein coding	-0.7549	3.38E-04	
<b>HERC3</b>	HECT and RLD domain containing E3 ubiquitin protein ligase 3	protein coding	-0.8287	3.62E-04	
<b>FHL1</b>	four and a half LIM domains 1	protein coding	-1.1199	3.75E-04	



IGDCC3	immunoglobulin superfamily DCC subclass member 3	protein coding	-0.8229	4.45E-04	
<i>DLEU1_2</i>	ENSG00000273541 Deleted in lymphocytic leukemia 1 conserved region 2	misc RNA	-3.7689	4.48E-04	
<b>LPCAT2</b>	lysophosphatidylcholine acyltransferase 2	protein coding	-1.0358	4.82E-04	
B4GALNT3	beta-1,4-N-acetyl-galactosaminyltransferase 3	protein coding	-1.0827	5.42E-04	
ANKRD35	ankyrin repeat domain 35	protein coding	-1.1008	5.70E-04	
KCNJ8	potassium voltage-gated channel subfamily J member 8	protein coding	-1.3980	5.75E-04	
OAS1	2'-5'-oligoadenylate synthetase 1	protein coding	-1.2567	5.77E-04	
<b>LAMB1</b>	laminin subunit beta 1	protein coding	-0.6224	6.04E-04	
<b>MATN3</b>	matrilin 3	protein coding	-1.9791	6.04E-04	
ADRA2C	adrenoceptor alpha 2C	protein coding	-0.6003	6.04E-04	
FST	folistatin	protein coding	-1.9604	6.05E-04	
LUZP2	leucine zipper protein 2	protein coding	-1.2511	6.20E-04	
<b>CASC10</b>	cancer susceptibility 10	protein coding	-0.8790	6.22E-04	
PCDH9	protocadherin 9	protein coding	-0.6212	6.51E-04	
BST2	bone marrow stromal cell antigen 2	protein coding	-1.8617	6.59E-04	
<b>TGFB2</b>	transforming growth factor beta 2	protein coding	-0.7966	6.74E-04	
<b>ANXA6</b>	annexin A6	protein coding	-0.5892	7.57E-04	
<b>MUM1L1</b>	MUM1 like 1	protein coding	-0.7705	7.58E-04	
LGALS1	galectin 1	protein coding	-0.8794	8.43E-04	
PDLIM7	PDZ and LIM domain 7	protein coding	-0.6039	8.46E-04	
IFITM1	interferon induced transmembrane protein 1	protein coding	-2.5758	8.80E-04	
IFI44L	interferon induced protein 44 like	protein coding	-2.6850	9.00E-04	
<i>AL590004.4</i>	ENSG00000260604	lincRNA	-1.1536	1.04E-03	
H19	H19, imprinted maternally expressed transcript (non-protein coding)	processed transcript	-1.0145	1.12E-03	
ARHGEF40	Rho guanine nucleotide exchange factor 40	protein coding	-1.0189	1.12E-03	
IFI27	interferon alpha inducible protein 27	protein coding	-2.4131	1.15E-03	
<b>ST6GALNAC2</b>	ST6 N-acetylgalactosaminide alpha-2,6-sialyltransferase 2	protein coding	-0.6832	1.21E-03	
IRF7	interferon regulatory factor 7	protein coding	-0.9153	1.40E-03	
MAATS1	MYCBP associated and testis expressed 1	protein coding	-2.2043	1.52E-03	
PLEKHG2	pleckstrin homology and RhoGEF domain containing G2	protein coding	-0.6605	1.56E-03	
TRGC1	T-cell receptor gamma constant 1	TR C gene	-1.0348	1.64E-03	
<b>DIO2</b>	iodothyronine deiodinase 2	protein coding	-0.8343	1.71E-03	
<b>SEPT10</b>	septin 10	protein coding	-0.5945	1.77E-03	
<b>KIAA1210</b>	KIAA1210	protein coding	-2.2030	1.85E-03	
<b>PIK3C2G</b>	phosphatidylinositol-4-phosphate 3-kinase catalytic subunit type 2 gamma	protein coding	-0.6057	1.87E-03	
PSG9	pregnancy specific beta-1-glycoprotein 9	protein coding	-1.6659	1.88E-03	
<i>PRICKLE2-AS1</i>	ENSG00000241111 PRICKLE2 antisense RNA 1	antisense	-0.7703	1.88E-03	
TIMP1	TIMP metalloproteinase inhibitor 1	protein coding	-0.6725	1.89E-03	
<b>GYG2</b>	glycogenin 2	protein coding	-0.7694	1.90E-03	
TGM2	transglutaminase 2	protein coding	-1.4303	1.94E-03	
RBM24	RNA binding motif protein 24	protein coding	-0.7566	2.05E-03	
<b>EMP1</b>	epithelial membrane protein 1	protein coding	-1.6130	2.12E-03	
ANTXR2	anthrax toxin receptor 2	protein coding	-0.9998	2.26E-03	
PSG4	pregnancy specific beta-1-glycoprotein 4	protein coding	-3.3967	2.26E-03	



<b>TCTN2</b>	tectonic family member 2	protein coding	-0.7736	2.26E-03	
<b>CLIP4</b>	CAP-Gly domain containing linker protein family member 4	protein coding	-1.0533	2.39E-03	
MTMR11	myotubularin related protein 11	protein coding	-0.8150	2.40E-03	
<b>MYPN</b>	myopalladin	protein coding	-1.5510	2.41E-03	
DDX60L	DEAD-box helicase 60 like	protein coding	-0.6835	2.45E-03	
HIST1H2BM	histone cluster 1 H2B family member m	protein coding	-0.9638	2.49E-03	
CALCR	calcitonin receptor	protein coding	-0.8435	2.72E-03	
KRT17	keratin 17	protein coding	-1.9283	2.74E-03	
SALL4	spalt like transcription factor 4	protein coding	-0.6768	2.90E-03	
<b>HCAR1</b>	hydroxycarboxylic acid receptor 1	protein coding	-0.6711	2.97E-03	
MUCL1	mucin like 1	protein coding	-0.6291	3.13E-03	
AREG	amphiregulin	protein coding	-1.3155	3.84E-03	
<b>LINC00472</b>	long intergenic non-protein coding RNA 472	lincRNA	-0.7794	3.91E-03	
CTGF	connective tissue growth factor	protein coding	-1.6429	4.13E-03	
LAMB3	laminin subunit beta 3	protein coding	-1.4995	4.18E-03	
BMPER	BMP binding endothelial regulator	protein coding	-1.1140	4.45E-03	
<b>CYR61</b>	cysteine rich angiogenic inducer 61	protein coding	-0.8379	4.46E-03	
ADGRF4	adhesion G protein-coupled receptor F4	protein coding	-1.8521	4.51E-03	
GHR	growth hormone receptor	protein coding	-0.7498	4.80E-03	
<b>TMPRSS11E</b>	transmembrane protease, serine 11E	protein coding	-1.6753	4.84E-03	
CCDC74B	coiled-coil domain containing 74B	protein coding	-1.2445	5.05E-03	
NTN1	netrin 1	protein coding	-0.9267	5.12E-03	
ADGRL2	adhesion G protein-coupled receptor L2	protein coding	-0.8940	5.13E-03	
SHH	sonic hedgehog	protein coding	-1.1061	5.26E-03	
CYBRD1	cytochrome b reductase 1	protein coding	-0.6508	5.39E-03	
TNFRSF19	TNF receptor superfamily member 19	protein coding	-1.4601	5.83E-03	
<b>LYN</b>	LYN proto-oncogene, Src family tyrosine kinase	protein coding	-1.0818	5.83E-03	
<b>ARSJ</b>	arylsulfatase family member J	protein coding	-0.6244	6.01E-03	
CEACAMP10	carcinoembryonic antigen related cell adhesion molecule pseudogene 10	pseudogene (TP)	-3.1026	6.16E-03	
GCNT4	glucosaminyl (N-acetyl) transferase 4, core 2	protein coding	-0.6922	6.19E-03	
GPOR1	G protein-coupled estrogen receptor 1	protein coding	-1.3619	6.42E-03	
<b>PTGER2</b>	prostaglandin E receptor 2	protein coding	-2.0768	6.76E-03	
HIST2H3A	histone cluster 2 H3 family member a	protein coding	-0.7627	7.09E-03	
GBP1	guanylate binding protein 1	protein coding	-1.6321	7.13E-03	
SLITRK5	SLIT and NTRK like family member 5	protein coding	-3.0023	7.13E-03	
<b>NECTIN3</b>	nectin cell adhesion molecule 3	protein coding	-0.7068	7.18E-03	
<b>AC009533.1</b>	ENSG00000111788	pseudogene (U)	-0.6127	7.23E-03	
ARL4D	ADP ribosylation factor like GTPase 4D	protein coding	-0.6564	7.60E-03	
CALD1	caldesmon 1	protein coding	-1.4531	7.68E-03	
<b>AL121603.2</b>	ENSG00000258738	antisense	-1.0886	7.77E-03	
SCN1B	sodium voltage-gated channel beta subunit 1	protein coding	-0.9294	7.87E-03	
TMCC3	transmembrane and coiled-coil domain family 3	protein coding	-1.3806	7.99E-03	
ZIC3	Zic family member 3	protein coding	-1.2911	8.03E-03	
ITM2A	integral membrane protein 2A	protein coding	-0.6929	8.05E-03	
GCHFR	GTP cyclohydrolase I feedback regulator	protein coding	-0.7220	8.46E-03	
SERPINE1	serpin family E member 1	protein coding	-1.4173	9.45E-03	

ERFE	erythroferrone	protein coding	-1.1041	9.68E-03	
ARHGAP22	Rho GTPase activating protein 22	protein coding	-1.1118	9.71E-03	
LINC02412*	long intergenic non-protein coding RNA 2412	lincRNA	-2.7792	9.75E-03	
AC090568.2	ENSG00000253553	antisense	-2.7316	9.95E-03	
ANKRD62P1	ankyrin repeat domain 62 pseudogene 1	pseudogene (TU)	-3.1856	0.0104	
PGM5	phosphoglucomutase 5	protein coding	-2.7654	0.0109	
MFGE8	milk fat globule-EGF factor 8 protein	protein coding	-0.8024	0.0115	
C15orf59	chromosome 15 open reading frame 59	protein coding	-1.1690	0.0115	
PLPPR5	phospholipid phosphatase related 5	protein coding	-0.7519	0.0117	
KLK11	kallikrein related peptidase 11	protein coding	-2.6484	0.0117	
AL390038.1	ENSG00000224698	lincRNA	-4.0293	0.0119	
APCDD1	APC down-regulated 1	protein coding	-0.7783	0.0120	
GPR137C	G protein-coupled receptor 137C	protein coding	-0.7389	0.0123	
TPM2	tropomyosin 2 (beta)	protein coding	-0.7226	0.0130	
AARD	alanine and arginine rich domain containing protein	protein coding	-0.7363	0.0135	
AC022113.1	ENSG00000246214	antisense	-2.5946	0.0144	
SLC8A1	solute carrier family 8 member A1	protein coding	-1.6197	0.0153	
AC126175.1	ENSG00000277738	lincRNA	-1.5596	0.0158	
TGFBR2	transforming growth factor beta receptor 2	protein coding	-0.6337	0.0159	
AFF2	AF4/FMR2 family member 2	protein coding	-1.5386	0.0164	
CLIC3	chloride intracellular channel 3	protein coding	-1.5111	0.0180	
AC021218.1	ENSG00000204876	lincRNA	-2.8117	0.0182	
DOK7	docking protein 7	protein coding	-0.9845	0.0187	
AC131097.4	ENSG00000235151	lincRNA	-2.5705	0.0192	
KRT14	keratin 14	protein coding	-2.7049	0.0198	
FOS	Fos proto-oncogene, AP-1 transcription factor subunit	protein coding	-0.7177	0.0199	
NCMAP	non-compact myelin associated protein	protein coding	-1.3002	0.0199	
OLFML2A	olfactomedin like 2A	protein coding	-1.1451	0.0209	
CHD5	chromodomain helicase DNA binding protein 5	protein coding	-0.6334	0.0211	
PAK6	p21 (RAC1) activated kinase 6	protein coding	-2.4392	0.0211	
MERTK	MER proto-oncogene, tyrosine kinase	protein coding	-0.9955	0.0213	
CALHM2	calcium homeostasis modulator 2	protein coding	-0.7397	0.0215	
NOV	nephroblastoma overexpressed	protein coding	-2.1548	0.0216	
AC092902.2	ENSG00000241288	processed transcript	-0.6634	0.0218	
TGIF2-C20orf24	TGIF2-C20orf24 readthrough	protein coding	-3.7336	0.0222	
PPP2R2B	protein phosphatase 2 regulatory subunit Bbeta	protein coding	-1.6917	0.0232	
PSG5	pregnancy specific beta-1-glycoprotein 5	protein coding	-2.5035	0.0234	
AC008264.2	ENSG00000273489	antisense	-2.4319	0.0235	
LINC00355	long intergenic non-protein coding RNA 355	lincRNA	-0.6426	0.0237	
MAP3K14	mitogen-activated protein kinase kinase kinase 14	protein coding	-0.6163	0.0241	
STON1	stonin 1	protein coding	-0.7562	0.0241	
SYTL5	synaptotagmin like 5	protein coding	-1.0902	0.0242	
AL513211.1	ENSG00000228614	antisense	-2.4706	0.0242	
KLHL4	kelch like family member 4	protein coding	-2.1559	0.0249	
DRP2	dystrophin related protein 2	protein coding	-1.1196	0.0249	
RGPD6	RANBP2-like and GRIP domain containing 6	protein coding	-0.7453	0.0250	

NPIPB1P	nuclear pore complex interacting protein family member B1, pseudogene	pseudogene (TU)	-2.5832	0.0250	
SLCO2A1	solute carrier organic anion transporter family member 2A1	protein coding	-1.1248	0.0252	
AC093001.1	ENSG00000244468	antisense	-1.0772	0.0253	
GLB1L	galactosidase beta 1 like	protein coding	-0.6830	0.0254	
CCL26	C-C motif chemokine ligand 26	protein coding	-1.8865	0.0256	
GPC5	glypican 5	protein coding	-1.2306	0.0258	
GP1BB	glycoprotein Ib platelet beta subunit	protein coding	-2.5674	0.0261	
AC010733.2	ENSG00000267520	3' overlapping ncRNA	-3.6008	0.0266	
SEMA5A	semaphorin 5A	protein coding	-0.8145	0.0268	
LRRC15	leucine rich repeat containing 15	protein coding	-1.6985	0.0269	
PTPRE	protein tyrosine phosphatase, receptor type E	protein coding	-1.0532	0.0273	
CITED1	Cbp/p300 interacting transactivator with Glu/Asp rich carboxy-terminal domain 1	protein coding	-1.5737	0.0274	
ITIH2	inter-alpha-trypsin inhibitor heavy chain 2	protein coding	-1.3829	0.0275	
CPA4	carboxypeptidase A4	protein coding	-2.2158	0.0279	
IL1RN	interleukin 1 receptor antagonist	protein coding	-2.3217	0.0284	
MAFB	MAF bZIP transcription factor B	protein coding	-0.6698	0.0285	
AC022360.1	ENSG00000253156	pseudogene (P)	-2.5833	0.0293	
AC092431.2	ENSG00000232228	pseudogene (P)	-2.3917	0.0297	
PTCH1	patched 1	protein coding	-0.7922	0.0299	
WISP2	WNT1 inducible signaling pathway protein 2	protein coding	-0.8951	0.0311	
ITGB3	integrin subunit beta 3	protein coding	-2.2754	0.0311	
AL136295.5	ENSG00000259529	protein coding	-0.6815	0.0316	
SPAM1	sperm adhesion molecule 1	protein coding	-2.3023	0.0320	
CAMK1	calcium/calmodulin dependent protein kinase I	protein coding	-0.7587	0.0324	
ERBB4	erb-b2 receptor tyrosine kinase 4	protein coding	-0.7505	0.0324	
TLE4	transducin like enhancer of split 4	protein coding	-1.3736	0.0325	
PLEKHA4	pleckstrin homology domain containing A4	protein coding	-1.7836	0.0330	
MXRA7	matrix remodeling associated 7	protein coding	-0.6761	0.0338	
CNIH2	cornichon family AMPA receptor auxiliary protein 2	protein coding	-0.6491	0.0341	
SEMA6B	semaphorin 6B	protein coding	-0.6062	0.0354	
CCDC74A	coiled-coil domain containing 74A	protein coding	-0.6788	0.0357	
ZACN	zinc activated ion channel	protein coding	-2.4210	0.0359	
KCNF1	potassium voltage-gated channel modifier subfamily F member 1	protein coding	-1.4974	0.0363	
GGTA1P	glycoprotein, alpha-galactosyltransferase 1 pseudogene	unitary pseudogene (T)	-2.0282	0.0364	
CASP14	caspase 14	protein coding	-1.5200	0.0365	
GALNT16	polypeptide N-acetylgalactosaminyltransferase 16	protein coding	-0.8723	0.0367	
AL031123.2	ENSG00000261211	lincRNA	-0.8757	0.0379	
UBE2QL1	ubiquitin conjugating enzyme E2 Q family like 1	protein coding	-2.3441	0.0386	
FOXF1	forkhead box F1	protein coding	-2.1948	0.0393	
SLX1A	SLX1 homolog A, structure-specific endonuclease subunit	protein coding	-0.6784	0.0395	
MUC5B	mucin 5B, oligomeric mucus/gel-forming	protein coding	-1.0043	0.0396	
MIR503HG	MIR503 host gene	lincRNA	-0.6001	0.0404	

<i>AC046143.1</i>	ENSG00000229334	antisense	-1.7197	0.0408	
C8orf4	chromosome 8 open reading frame 4	protein coding	-1.2589	0.0412	
IGFBP6	insulin like growth factor binding protein 6	protein coding	-0.8275	0.0417	
BMPR1B-AS1	BMPR1B antisense RNA 1 (head to head)	lincRNA	-2.2770	0.0417	
NPM1P19	nucleophosmin 1 pseudogene 19	pseudogene (P)	-2.2135	0.0425	
HLA-DPA1	major histocompatibility complex, class II, DP alpha 1	protein coding	-2.4023	0.0425	
A2M-AS1	A2M antisense RNA 1 (head to head)	antisense	-1.0127	0.0434	
NES	nestin	protein coding	-0.6837	0.0435	
HNRNPKP2	heterogeneous nuclear ribonucleoprotein K pseudogene 2	pseudogene (P)	-2.2768	0.0436	
DDX11L10	DEAD/H-box helicase 11 like 10	pseudogene (TU)	-2.3020	0.0440	
GPR85	G protein-coupled receptor 85	protein coding	-1.7987	0.0442	
WEE2	WEE1 homolog 2	protein coding	-2.1995	0.0444	
SGCB	sarcoglycan beta	protein coding	-0.6789	0.0447	
SAP25	Sin3A associated protein 25	protein coding	-2.3072	0.0447	
ZNF365	zinc finger protein 365	protein coding	-0.7436	0.0458	
AOX1	aldehyde oxidase 1	protein coding	-0.7992	0.0461	
CMPK2	cytidine/uridine monophosphate kinase 2	protein coding	-1.4018	0.0465	
CTNNA3	catenin alpha 3	protein coding	-2.1505	0.0473	
<b>PALMD</b>	palmdelphin	protein coding	-0.6745	0.0477	
NPAS3	neuronal PAS domain protein 3	protein coding	-0.6731	0.0483	
ATP6V1G2-DDX39B	ATP6V1G2-DDX39B readthrough (NMD candidate)	protein coding	-0.7292	0.0484	
PLAU	plasminogen activator, urokinase	protein coding	-0.6551	0.0492	
CDH12	cadherin 12	protein coding	-0.6140	0.0493	
PDE4A	phosphodiesterase 4A	protein coding	-0.9036	0.0493	
ACOX2	acyl-CoA oxidase 2	protein coding	-1.3430	0.0494	
<i>AC106028.5</i>	ENSG00000275175	lincRNA	-2.3021	0.0494	

**Additional Table 4.2.** List of 291 genes down-regulated following ELF5 induction, with absolute fold change >1.5 and FDR <0.05, sorted by FDR. “ChIP” column indicates the presence (purple) or absence (white) of an ELF5 ChIP-seq peak within 10kb upstream of the TSS. Genes in blue typeface (total 88) were also found to be down-regulated in the MCF7-ELF5-V5 microarray experiment. Gene names in italics indicate they have no official HGNC gene symbol; Ensembl gene names and identifiers (beginning ‘ENSG’) are provided for these genes. One gene (official symbol LINC02412) is marked with an asterisk (\*), indicating that it did not have an HGNC symbol when the data were originally analysed but has since been updated. P, processed; TP, transcribed processed; TU, transcribed unprocessed; U, unprocessed.

### Additional Tables 4.3: Heatmap gene symbols and log2 fold change values

#### A Dutertre Estradiol Response 24hr Up

(Figure 4.6A)

Gene symbol	RNA-seq Log2fc	Microarray Log2fc
AREG	-1.3155	-0.3999
SYTL5	-1.0902	-0.4907
H19	-1.0145	-0.5127
WISP2	-0.8951	-0.0707
COL12A1	-0.8603	-1.1207
PCP4	-0.7357	0.0180
KRT13	-0.7025	0.0239
CNIH2	-0.6491	-0.4169
NFATC2	-0.6118	0.0988
STC2	-0.5667	-0.5134
ELOVL2	-0.5595	-0.8761
FAM111B	-0.5033	-0.7775
LYPD6	-0.4973	-0.6607
CHPT1	-0.4953	-0.3598
SLC26A2	-0.4799	-0.1280
FANCD2	-0.4690	-0.6592
CCNE2	-0.4642	-0.5383
ADCY1	-0.4599	-0.2417
IFITM10	-0.4484	NaN
MYBL1	-0.4410	-0.5219
SPC24	-0.4371	-0.5156
CENPI	-0.4298	-0.7822
RAPGEFL1	-0.4194	-0.5973
E2F8	-0.4030	-0.5558
RET	-0.4028	-0.3223
KCNK15	-0.3965	0.0418
CDK1	-0.3926	-0.4626
RBL1	-0.3876	-0.1331
ANLN	-0.3822	-0.5837
FEN1	-0.3809	-0.2333
FAM83D	-0.3790	-0.2605
RAD54B	-0.3783	-0.4963
CHAF1A	-0.3746	-0.3370
RAD51	-0.3743	-0.5026
PBK	-0.3711	-0.4766
BRIP1	-0.3626	-0.4399
ASPM	-0.3612	-0.5065
C21ORF58	-0.3594	-0.2707
DTL	-0.3558	-0.5223
CDC45	-0.3550	-0.3250
POLD3	-0.3550	-0.4421
KIAA1524	-1.3155	-0.3999

Gene symbol	RNA-seq Log2fc	Microarray Log2fc
KIAA1524	-0.3547	-0.4045
TET2	-0.3544	-0.4132
ARL3	-0.3541	-0.1454
HIST1H1C	-0.3516	-0.4118
STC1	-0.3452	-0.5076
SLC39A8	-0.3439	-0.5000
MCM4	-0.3432	-0.4038
PKIB	-0.3411	-0.1692
ESCO2	-0.3390	-0.5340
BUB1	-0.3378	-0.5135
CHTF18	-0.3370	-0.4048
GIN51	-0.3366	-0.4159
POLA2	-0.3355	-0.4575
CLSPN	-0.3353	-0.2685
KIF23	-0.3298	-0.4831
HELLS	-0.3289	-0.5472
UHRF1	-0.3260	-0.5270
GIN52	-0.3205	-0.2627
KIF11	-0.3195	-0.3783
MCM10	-0.3193	-0.4959
UBE2C	-0.3184	-0.1329
WDR76	-0.3173	-0.5647
MCM7	-0.3171	-0.3176
MASTL	-0.3167	-0.4025
ATAD2	-0.3162	-0.3393
POLQ	-0.3137	-0.5172
MCM3	-0.3122	-0.3967
RAD54L	-0.3121	-0.5269
CDCA7	-0.3110	-0.6903
FANCI	-0.3065	-0.4324
NCAPG	-0.3063	-0.4159
NCAPH	-0.3060	-0.2913
TOP2A	-0.3060	-0.4637
GIN53	-0.3051	-0.2571
ARHGAP11A	-0.3035	-0.3783
PSMC3IP	-0.3027	-0.2484
KIFC1	-0.3008	-0.4260
HIST1H1E	-0.2979	-0.1873
FIGNL1	-0.2964	-0.4938
AURKA	-0.2962	-0.3566
CDK2	-0.2956	-0.3313
DEPDC1B	-0.2954	-0.3703

Gene symbol	RNA-seq Log2fc	Microarray Log2fc
RRM1	-0.2950	-0.3604
DLGAP5	-0.2930	-0.4029
RFC5	-0.2926	-0.2898
FBXO5	-0.2916	-0.2968
ZWILCH	-0.2915	-0.4498
BCL2	-0.2898	-0.3738
CHAF1B	-0.2878	-0.4168
E2F2	-0.2859	-0.4043
SMC2	-0.2767	-0.3269
ZNF367	-0.2760	-0.3720
CENPM	-0.2758	-0.0834
LIG1	-0.2756	-0.4101
STMN1	-0.2753	-0.1461
CCNA2	-0.2740	-0.3452
MCM5	-0.2724	-0.4094
ESPL1	-0.2718	-0.4484
MCM6	-0.2717	-0.4352
WDR34	-0.2711	-0.2566
TK1	-0.2694	-0.2173
PAQR4	-0.2681	-0.3898
XRCC2	-0.2681	-0.5730
BLM	-0.2675	-0.3333
RFC4	-0.2659	-0.3138
FANCC	-0.2637	-0.4217
ASF1B	-0.2633	-0.2663
MELK	-0.2631	-0.3725
CELSR2	-0.2628	-0.1864
TROAP	-0.2625	-0.5118
PRC1	-0.2624	-0.3930
PLK4	-0.2616	-0.4285
RPA3	-0.2584	-0.0146
PCNA	-0.2571	-0.2531
KCNK5	-0.2561	-0.4572
DSCC1	-0.2560	-0.3618
CDC6	-0.2558	-0.2611
RRM2	-0.2539	-0.1526
CBX5	-0.2502	-0.3465
CCNB2	-0.2496	-0.2115
ATAD5	-0.2474	-0.5549
FKBP5	-0.2453	-0.3913
WDHD1	-0.2444	-0.4361
PTTG1	-0.2441	-0.0690
AURKB	-0.2439	-0.2279
GMNN	-0.2439	-0.1894
BRCA2	-0.2418	-0.2257
EBP	-0.2418	-0.2256

Gene symbol	RNA-seq Log2fc	Microarray Log2fc
RFC2	-0.2410	-0.2402
SHCBP1	-0.2386	-0.3035
MCM8	-0.2382	-0.4239
TRAIP	-0.2368	-0.2301
FANCG	-0.2336	-0.2760
SPAG5	-0.2331	-0.5516
RAB31	-0.2324	-0.2000
MMS22L	-0.2296	-0.2103
IQGAP3	-0.2275	-0.2807
TRIP13	-0.2264	-0.4449
KIF2C	-0.2262	-0.1375
KNTC1	-0.2256	-0.4730
MYBL2	-0.2247	-0.2292
TYMS	-0.2243	-0.3204
PKMYT1	-0.2221	-0.3556
BIRC5	-0.2221	-0.1942
POLD1	-0.2211	-0.2408
LRR1	-0.2205	NaN
TMPO	-0.2205	-0.1883
CDCA2	-0.2191	-0.3723
CHEK1	-0.2167	-0.2062
POLE	-0.2163	-0.3474
IL17RB	-0.2141	-0.5512
NUSAP1	-0.2140	-0.3577
SULF1	-0.2139	-0.7373
GGH	-0.2127	0.0592
BRCA1	-0.2109	-0.2865
MICB	-0.2104	-0.0877
ZWINT	-0.2096	-0.5769
DNA2	-0.2096	-0.4787
CIT	-0.2070	-0.4683
MIS18A	-0.2065	NaN
TTK	-0.2057	-0.3925
RNASEH2A	-0.2048	-0.2076
NUP107	-0.2034	-0.2484
HAUS4	-0.2028	-0.1929
H2AFX	-0.2028	-0.1879
EXO1	-0.2022	-0.3748
NCAPG2	-0.2018	-0.3264
GAS2L3	-0.1994	-0.3826
FANCA	-0.1987	-0.4934
IL1RAP	-0.1976	-0.0958
DUT	-0.1969	-0.2221
MPHOSPH9	-0.1955	-0.3722
STIL	-0.1954	-0.2810
UBE2T	-0.1940	-0.1842

Gene symbol	RNA-seq Log2fc	Microarray Log2fc
CDCA4	-0.1929	-0.1721
CENPO	-0.1921	0.2081
MAP6D1	-0.1902	-0.1452
MCM2	-0.1898	-0.3292
DCLRE1B	-0.1890	-0.1074
SUV39H1	-0.1848	-0.1211
RECQL4	-0.1842	-0.2358
TPX2	-0.1840	-0.3358
NUP85	-0.1834	-0.2626
TCF19	-0.1833	-0.3412
MAN1A1	-0.1819	-0.4351
TONSL	-0.1819	NaN
RFC3	-0.1808	-0.0557
NRIP1	-0.1793	-0.0335
JAK2	-0.1771	-0.1811
UNG	-0.1760	-0.1975
NCAPD3	-0.1743	-0.3634
TIMELESS	-0.1736	-0.3301
FAM46C	-0.1728	-0.1509
GFRA1	-0.1679	-0.1140
E2F7	-0.1679	-0.3899
CENPN	-0.1662	-0.0028
CENPL	-0.1653	-0.1287
TACC3	-0.1611	-0.3440
LMNB1	-0.1579	-0.4214
DSN1	-0.1577	-0.3701
NASP	-0.1577	-0.2769
BUB1B	-0.1548	-0.3848
CEP55	-0.1531	-0.3828
SNRNP25	-0.1523	-0.1681
DSCAM	-0.1501	-0.4277
DNMT1	-0.1457	-0.4635
CEP78	-0.1455	-0.2661
KIF4A	-0.1434	-0.3927
TREX2	-0.1433	-0.0212
CHRNA5	-0.1427	-0.2477
RMI1	-0.1422	-0.2758
HAUS8	-0.1412	-0.1031
MTHFD1	-0.1406	-0.3384
CXCL12	-0.1385	-0.8102
PLK1	-0.1373	-0.2386
RACGAP1	-0.1340	-0.2460
NCAPD2	-0.1322	-0.2160
POLE2	-0.1320	-0.2730
TMED8	-0.1311	-0.0801
CCDC34	-0.1305	-0.1143

Gene symbol	RNA-seq Log2fc	Microarray Log2fc
CA12	-0.1295	-0.3405
SKP2	-0.1282	-0.3570
INCENP	-0.1277	-0.2457
VRK1	-0.1268	-0.2020
NCAPH2	-0.1255	-0.0683
SKA3	-0.1232	-0.2347
CDCA5	-0.1229	-0.0187
WDR62	-0.1220	-0.2847
MKI67	-0.1205	-0.3213
DHTKD1	-0.1135	-0.2346
CENPJ	-0.1110	-0.3687
SUV39H2	-0.1100	-0.1187
E2F1	-0.1093	-0.2622
BRI3BP	-0.1089	-0.2041
NOS1AP	-0.1082	-0.1083
POLA1	-0.1080	-0.2771
TMEM164	-0.1053	-0.3650
RAD18	-0.0996	-0.0383
NR2C2AP	-0.0963	0.0474
TTF2	-0.0862	-0.3538
TMEM38B	-0.0855	-0.0234
TFDP1	-0.0845	-0.3616
SLC22A5	-0.0811	-0.1810
GTSE1	-0.0799	-0.3028
PRIM1	-0.0756	-0.0912
FOXM1	-0.0660	-0.1800
MYO19	-0.0638	-0.1435
SLC2A1	-0.0571	0.0837
SLC9A3R1	-0.0567	-0.2644
RBBP8	-0.0564	-0.0385
FKBP4	-0.0538	-0.0444
HR	-0.0531	-0.1005
SFXN2	-0.0524	-0.3288
XRCC3	-0.0454	0.0677
POLD2	-0.0439	-0.1000
PRR11	-0.0421	-0.3231
SIAH2	-0.0379	0.0759
CEP85	-0.0334	NaN
JPH1	-0.0267	-0.3923
SGK3	-0.0234	0.1328
SEMA3B	-0.0183	-0.1365
TPD52L1	-0.0173	-0.1058
NPY1R	0.0022	-0.4839
DARS2	0.0069	-0.1260
IMPA2	0.0083	0.0162
NME1	0.0096	0.1491

Gene symbol	RNA-seq Log2fc	Microarray Log2fc
PGR	0.0120	0.0032
MANEAL	0.0192	0.0534
CDCA7L	0.0242	-0.0968
CDT1	0.0265	-0.0023
CTSD	0.0305	0.1142
EXOSC2	0.0309	-0.1241
FREM2	0.0354	-0.1070
AMZ1	0.0403	0.0982
C1QTNF6	0.0437	0.0110
GREB1	0.0501	-0.1375
PDSS1	0.0577	-0.0434
DDX10	0.0607	-0.0582
HPRT1	0.0609	0.2330
SLC39A6	0.0745	0.0486
L2HGDH	0.0757	0.2110
TMEM120B	0.0765	-0.3409
BARD1	0.0804	-0.1637
LSM4	0.0874	0.1224
EPS15L1	0.1133	-0.1003
ABHD2	0.1141	0.0262
PPIF	0.1265	-0.0307
IGFBP4	0.1356	0.0789
XYLB	0.1837	-0.0669
GLB1L2	0.1866	0.1876
EXOSC5	0.1974	0.1879
TFAP4	0.2111	-0.1914
SLC29A1	0.2342	-0.0048
CD320	0.2598	0.0602
TST	0.2916	0.5122
XBP1	0.3094	0.1135
SLC27A2	0.3365	0.3957
LRIG1	0.3586	0.0984
NXNL2	0.3589	-0.0982
SLC7A5	0.4168	0.2234
ARMCX6	0.4483	0.0468
RERG	0.5504	0.6159
MGP	0.6221	0.7903
KCNK6	0.6436	0.3545
TFF1	0.6745	0.6324
MYB	0.6876	0.0387
DHRS2	0.9587	1.2313
DEGS2	1.1826	-0.1053
GLA	1.2513	0.8888
DEPTOR	1.2595	NaN

**B Kobayashi EGFR Signaling 24hr Down (Figure 4.6A)**

Gene symbol	RNA-seq Log2fc	Microarray Log2fc
EREG	-1.7187	-0.1399
OASL	-1.5814	-0.3098
AXL	-1.5248	-0.156
DKK1	-1.3895	-1.1578
ACOX2	-1.3430	-0.0682
AREG	-1.3155	-0.3999
ETV5	-1.0779	-0.1883
HMGA2	-1.0478	-0.0583
IL11	-0.9510	0.043
GJB3	-0.7724	-0.1173
NT5E	-0.5846	-0.2969
HJURP	-0.5239	-0.5772
TNFRSF12A	-0.5151	-0.358
DUSP6	-0.4907	0.2056
MAFF	-0.4814	-0.0615
CCNE2	-0.4642	-0.5383
MYBL1	-0.4410	-0.5219
UPP1	-0.4138	-0.1038
E2F8	-0.4030	-0.5558
CDK1	-0.3926	-0.4626
SPC25	-0.3819	-0.3495
FEN1	-0.3809	-0.2333
RAD54B	-0.3783	-0.4963
CHAF1A	-0.3746	-0.337
RAD51	-0.3743	-0.5026
PBK	-0.3711	-0.4766
DHFR	-0.3676	-0.3372
RAD51AP1	-0.3660	-0.4287
ASPM	-0.3612	-0.5065
CX3CL1	-0.3591	-0.0998
DTL	-0.3558	-0.5223
CDC45	-0.3550	-0.325
HAT1	-0.3462	-0.1254
STC1	-0.3452	-0.5076
MCM4	-0.3432	-0.4038
BUB1	-0.3378	-0.5135
GIN51	-0.3366	-0.4159
POLA2	-0.3355	-0.4575
KIF23	-0.3298	-0.4831
HELLS	-0.3289	-0.5472
TUBB4B	-0.3227	NaN



Gene symbol	RNA-seq Log2fc	Microarray Log2fc
GINS2	-0.3205	-0.2627
KIF11	-0.3195	-0.3783
MCM10	-0.3193	-0.4959
UBE2C	-0.3184	-0.1329
MCM7	-0.3171	-0.3176
ATAD2	-0.3162	-0.3393
PSRC1	-0.3142	-0.6342
MCM3	-0.3122	-0.3967
ORC1	-0.3096	-0.2549
FANCI	-0.3065	-0.4324
NCAPG	-0.3063	-0.4159
NCAPH	-0.3060	-0.2913
TOP2A	-0.3060	-0.4637
GINS3	-0.3051	-0.2571
PSMC3IP	-0.3027	-0.2484
TUBB2A	-0.3005	-0.0943
CCNB1	-0.2985	-0.2346
AURKA	-0.2962	-0.3566
CDK2	-0.2956	-0.3313
RRM1	-0.2950	-0.3604
DLGAP5	-0.2930	-0.4029
RFC5	-0.2926	-0.2898
ZWILCH	-0.2915	-0.4498
MET	-0.2857	-0.6319
CDC20	-0.2853	-0.2255
HMGB2	-0.2805	-0.3579
TUBB	-0.2783	-0.2498
KIF14	-0.2773	-0.5409
NEK2	-0.2771	-0.4587
SMC2	-0.2767	-0.3269
CENPM	-0.2758	-0.0834
CCNA2	-0.2740	-0.3452
NDC80	-0.2731	-0.33
MSH2	-0.2727	-0.3611
MCM5	-0.2724	-0.4094
ESPL1	-0.2718	-0.4484
MCM6	-0.2717	-0.4352
CCND1	-0.2699	0.0189
TK1	-0.2694	-0.2173
CARD10	-0.2685	-0.0313
BLM	-0.2675	-0.3333
RFC4	-0.2659	-0.3138
HIST1H4C	-0.2650	-0.2784
ASF1B	-0.2633	-0.2663
TFPI2	-0.2633	-0.1836
MELK	-0.2631	-0.3725

Gene symbol	RNA-seq Log2fc	Microarray Log2fc
PRC1	-0.2624	-0.393
PLK4	-0.2616	-0.4285
PCNA	-0.2571	-0.2531
CDCA3	-0.2570	-0.5403
ORC6	-0.2566	0.0658
DSCC1	-0.2560	-0.3618
CDC6	-0.2558	-0.2611
DUSP5	-0.2547	-0.1604
RRM2	-0.2539	-0.1526
AURKB	-0.2439	-0.2279
GMNN	-0.2439	-0.1894
SPHK1	-0.2422	-0.0133
FAM64A	-0.2412	-0.1952
RFC2	-0.2410	-0.2402
DNAJC9	-0.2402	0.095
KPNA2	-0.2392	-0.1718
SHCBP1	-0.2386	-0.3035
SMC4	-0.2361	-0.2726
CENPA	-0.2347	-0.3108
SPAG5	-0.2331	-0.5516
TUBB6	-0.2292	-0.1958
TRIP13	-0.2264	-0.4449
KIF2C	-0.2262	-0.1375
CDCA8	-0.2256	-0.2681
MYBL2	-0.2247	-0.2292
TYMS	-0.2243	-0.3204
PKMYT1	-0.2221	-0.3556
BIRC5	-0.2221	-0.1942
TMPO	-0.2205	-0.1883
CDC25A	-0.2189	-0.3466
NUSAP1	-0.2140	-0.3577
ERCC6L	-0.2133	-0.488
BRCA1	-0.2109	-0.2865
ZWINT	-0.2096	-0.5769
ENO2	-0.2071	-0.384
TTK	-0.2057	-0.3925
RNASEH2A	-0.2048	-0.2076
ECT2	-0.2034	-0.3287
H2AFX	-0.2028	-0.1879
EXO1	-0.2022	-0.3748
NCAPG2	-0.2018	-0.3264
KIF18B	-0.2010	-0.1795
DONSON	-0.2007	-0.3685
MSH6	-0.1982	-0.2504
IER3	-0.1977	-0.1094
DUT	-0.1969	-0.2221

Gene symbol	RNA-seq Log2fc	Microarray Log2fc
STIL	-0.1954	-0.281
GALNT10	-0.1943	-0.2789
CKS2	-0.1929	0.033
CDCA4	-0.1929	-0.1721
SNRPD1	-0.1921	0.0359
CCNF	-0.1899	-0.329
MCM2	-0.1898	-0.3292
RFWD3	-0.1896	-0.1263
TPX2	-0.1840	-0.3358
RFC3	-0.1808	-0.0557
TUBG1	-0.1770	-0.2862
UNG	-0.1760	-0.1975
NCAPD3	-0.1743	-0.3634
TIMELESS	-0.1736	-0.3301
SPRED2	-0.1697	-0.1685
SRSF7	-0.1686	-0.4551
CENPN	-0.1662	-0.0028
GPSM2	-0.1659	-0.4102
TUBA1B	-0.1611	-0.0259
MAD2L1	-0.1598	-0.0961
KIF15	-0.1589	-0.3544
LMNB1	-0.1579	-0.4214
DSN1	-0.1577	-0.3701
BUB1B	-0.1548	-0.3848
CEP55	-0.1531	-0.3828
DDX39A	-0.1514	NaN
USP1	-0.1507	-0.0721
KIF4A	-0.1434	-0.3927
CDKN3	-0.1432	-0.1865
RMI1	-0.1422	-0.2758
UBE2S	-0.1401	0.0182
PLK1	-0.1373	-0.2386
LMNB2	-0.1371	-0.1709
DEPDC1	-0.1348	-0.0834
RACGAP1	-0.1340	-0.246
POLE2	-0.1320	-0.273
CSE1L	-0.1319	-0.2014
DCBLD2	-0.1285	-0.247
RANBP1	-0.1271	-0.1412
VRK1	-0.1268	-0.202
MKI67	-0.1205	-0.3213
ZC3HAV1	-0.1156	-0.0127
E2F1	-0.1093	-0.2622
CKLF	-0.1085	NaN
POLA1	-0.1080	-0.2771
FAM111A	-0.1078	-0.2535

Gene symbol	RNA-seq Log2fc	Microarray Log2fc
HMMR	-0.1062	-0.2891
ODC1	-0.1043	0.0229
EZH2	-0.0971	-0.1775
PHLDA2	-0.0918	0.0182
TCOF1	-0.0846	-0.1146
TFDP1	-0.0845	-0.3616
HSPA14	-0.0833	0.0535
H2AFZ	-0.0832	0.0557
TIPIN	-0.0825	-0.1592
ELL2	-0.0814	0.0031
GTSE1	-0.0799	-0.3028
PRIM1	-0.0756	-0.0912
HAUS7	-0.0719	-0.0591
EEF1E1	-0.0707	0.0958
NETO2	-0.0684	-0.154
NUDT15	-0.0609	-0.0324
EXOSC8	-0.0559	0.1083
NOC3L	-0.0486	-0.0356
ADORA2B	-0.0486	-0.2445
NAA15	-0.0430	-0.0369
TMEM158	-0.0409	-0.0789
PNN	-0.0379	-0.4997
PAICS	-0.0303	-0.0829
RRP15	-0.0230	-0.0453
CKS1B	-0.0205	-0.1262
TGFA	-0.0149	0.026
CCND3	-0.0134	0.1742
PA2G4	0.0033	-0.0852
ABCE1	0.0051	0.0844
NME1	0.0096	0.1491
NOP56	0.0210	-0.2736
DBF4	0.0211	-0.125
CDT1	0.0265	-0.0023
VWTR1	0.0283	-0.1472
SNRPA1	0.0363	-0.7736
SLC20A1	0.0427	-0.1066
PARP2	0.0619	0.0481
POLR2D	0.0747	0.0869
SLCO4A1	0.0795	0.0187
BARD1	0.0804	-0.1637
FABP5	0.0821	0.6265
PUS7	0.0949	-0.0336
RANGAP1	0.0955	-0.1056
PRPS1	0.1034	-0.1467
SRM	0.1073	0.079
SOX9	0.1098	-0.2365

Gene symbol	RNA-seq Log2fc	Microarray Log2fc
UCK2	0.1237	0.0839
NOLC1	0.1249	0.1781
PPIF	0.1265	-0.0307
ITGA6	0.1684	0.0921
POLR3K	0.1927	0.2871
SLC29A1	0.2342	-0.0048
DUSP4	0.4631	0.3549
CDC42EP1	0.4742	0.2259
FAM86B1	0.5242	0.1178
SLC43A3	0.5272	0.2108
STEAP1	0.5458	0.3617
ETV1	0.5687	-0.3004
TCN1	0.5802	0.0803
PHLDA1	0.5874	-0.1619
FOSL1	0.7306	-0.0059
PSAT1	1.1827	0.3313

**C Rosty Cervical Cancer Proliferation  
Cluster (Figure 4.6A)**

Gene symbol	RNA-seq Log2fc	Microarray Log2fc
HJURP	-0.5239	-0.5772
CDKN2A	-0.4850	0.1603
CCNE2	-0.4642	-0.5383
E2F8	-0.4030	-0.5558
CDK1	-0.3926	-0.4626
FEN1	-0.3809	-0.2333
HSPB11	-0.3761	-0.3986
PBK	-0.3711	-0.4766
DHFR	-0.3676	-0.3372
RAD51AP1	-0.3660	-0.4287
ASPM	-0.3612	-0.5065
DTL	-0.3558	-0.5223
MCM4	-0.3432	-0.4038
KIF20A	-0.3419	-0.5213
BUB1	-0.3378	-0.5135
GINS1	-0.3366	-0.4159
POLA2	-0.3355	-0.4575
KIF23	-0.3298	-0.4831
HELLS	-0.3289	-0.5472
KIF20B	-0.3203	-0.488
KIF11	-0.3195	-0.3783
MCM10	-0.3193	-0.4959
UBE2C	-0.3184	-0.1329
ATAD2	-0.3162	-0.3393

Gene symbol	RNA-seq Log2fc	Microarray Log2fc
POLQ	-0.3137	-0.5172
FANCI	-0.3065	-0.4324
NCAPG	-0.3063	-0.4159
NCAPH	-0.3060	-0.2913
TOP2A	-0.3060	-0.4637
KIFC1	-0.3008	-0.426
CCNB1	-0.2985	-0.2346
AURKA	-0.2962	-0.3566
TMSB10	-0.2934	-0.0722
DLGAP5	-0.2930	-0.4029
FBXO5	-0.2916	-0.2968
CHAF1B	-0.2878	-0.4168
CDC20	-0.2853	-0.2255
HMGB2	-0.2805	-0.3579
KIF14	-0.2773	-0.5409
NEK2	-0.2771	-0.4587
SMC2	-0.2767	-0.3269
CENPM	-0.2758	-0.0834
CCNA2	-0.2740	-0.3452
NDC80	-0.2731	-0.33
ESPL1	-0.2718	-0.4484
TK1	-0.2694	-0.2173
PAQR4	-0.2681	-0.3898
ASF1B	-0.2633	-0.2663
MELK	-0.2631	-0.3725
PRC1	-0.2624	-0.393
RPA3	-0.2584	-0.0146
PCNA	-0.2571	-0.2531
CDCA3	-0.2570	-0.5403
CDC6	-0.2558	-0.2611
CENPF	-0.2542	-0.4547
RRM2	-0.2539	-0.1526
CCNB2	-0.2496	-0.2115
PTTG1	-0.2441	-0.069
AURKB	-0.2439	-0.2279
GMNN	-0.2439	-0.1894
EBP	-0.2418	-0.2256
KPNA2	-0.2392	-0.1718
SHCBP1	-0.2386	-0.3035
SMC4	-0.2361	-0.2726
CENPA	-0.2347	-0.3108
SPAG5	-0.2331	-0.5516
CENPE	-0.2288	-0.4233
TRIP13	-0.2264	-0.4449
KIF2C	-0.2262	-0.1375
CDCA8	-0.2256	-0.2681

Gene symbol	RNA-seq Log2fc	Microarray Log2fc
MYBL2	-0.2247	-0.2292
TYMS	-0.2243	-0.3204
BIRC5	-0.2221	-0.1942
TMPO	-0.2205	-0.1883
CHEK1	-0.2167	-0.2062
NUSAP1	-0.2140	-0.3577
ERCC6L	-0.2133	-0.488
GGH	-0.2127	0.0592
BRCA1	-0.2109	-0.2865
ZWINT	-0.2096	-0.5769
DNA2	-0.2096	-0.4787
TTK	-0.2057	-0.3925
ECT2	-0.2034	-0.3287
H2AFX	-0.2028	-0.1879
KIF18B	-0.2010	-0.1795
PAFAH1B3	-0.1961	0.0384
STIL	-0.1954	-0.281
CKS2	-0.1929	0.033
CCNF	-0.1899	-0.329
MCM2	-0.1898	-0.3292
TPX2	-0.1840	-0.3358
DNMT3B	-0.1781	-0.6047
PLEK2	-0.1687	-0.0732
TACC3	-0.1611	-0.344
MAD2L1	-0.1598	-0.0961
KIF15	-0.1589	-0.3544
LMNB1	-0.1579	-0.4214
BUB1B	-0.1548	-0.3848
CEP55	-0.1531	-0.3828
ANP32E	-0.1519	-0.3194
KIF4A	-0.1434	-0.3927
UBE2S	-0.1401	0.0182
PLK1	-0.1373	-0.2386
RACGAP1	-0.1340	-0.246
MKI67	-0.1205	-0.3213
COMMD8	-0.1197	0.0277
APOBEC3B	-0.1128	-0.5831
HN1	-0.1103	0.0503
E2F1	-0.1093	-0.2622
HMMR	-0.1062	-0.2891
EZH2	-0.0971	-0.1775
DSG2	-0.0867	-0.0313
H2AFZ	-0.0832	0.0557
GTSE1	-0.0799	-0.3028
GIN54	-0.0791	-0.2729
NETO2	-0.0684	-0.154

Gene symbol	RNA-seq Log2fc	Microarray Log2fc
FOXMI	-0.0660	-0.18
BID	-0.0478	0.0824
DPP3	-0.0414	-0.1323
ACACA	-0.0169	-0.1658
DTYMK	0.0076	0.0986
HMGAI	0.0127	-0.0429
SAC3D1	0.0200	0.077
DBF4	0.0211	-0.125
OIP5	0.0594	0.0388
SLC38A1	0.0636	-0.0134
LRP8	0.0773	-0.1565
SLC25A15	0.0778	-0.1827
LSM4	0.0874	0.1224
MRPS15	0.1205	0.0773
MAPK13	0.1998	-0.0051
CELSR3	0.2927	0.0714
EIF4EBP1	0.6551	0.6136
CA2	1.6530	-0.0088

#### D Reactome 3' UTR Mediated Translational Regulation (Figure 4.6B)

Gene symbol	RNA-seq Log2fc	Microarray Log2fc
FAM153A	-1.0077	NaN
RPS19P3	-1.0003	NaN
RPS29P3	-0.8955	NaN
RPL7AP66	-0.4292	NaN
RPL21P75	-0.3606	NaN
RPL17P36	-0.2213	NaN
RPL5P1	-0.2151	NaN
EIF4E	-0.1859	-0.1330
RPL23AP18	-0.1821	NaN
RPS28P7	-0.1103	NaN
RPS26	-0.0691	0.0180
EIF1AX	-0.0601	-0.0048
RPL41	-0.0534	-0.0004
RPL3L	-0.0528	0.0631
RPS26P28	-0.0505	NaN
RPL21P16	-0.0482	NaN
EIF4H	-0.0409	-0.0577
RPL38	-0.0309	0.2651
EIF4G1	-0.0183	-0.1292
RPS15	0.0000	0.0384
EIF2S1	0.0029	-0.0037
EIF3I	0.0039	-0.0487

Gene symbol	RNA-seq Log2fc	Microarray Log2fc
EIF4A1	0.0041	-0.0019
EIF2S3	0.0054	-0.1099
EIF3B	0.0057	-0.0444
RPL35A	0.0093	0.1595
RPL26L1	0.0111	-0.0392
EIF3K	0.0160	0.0873
RPL23AP42	0.0209	NaN
RPL29	0.0232	0.1424
RPL23AP2	0.0304	NaN
RPL37A	0.0330	0.0947
UBA52	0.0366	0.1356
EIF3FP3	0.0376	NaN
RPL23AP63	0.0458	NaN
RPL15	0.0484	0.0491
EIF3H	0.0535	0.0636
RPLP1	0.0540	0.2898
RPS27AP11	0.0564	NaN
RPSA	0.0565	0.1233
RPS10	0.0580	0.3682
RPS12	0.0588	0.3266
RPL23A	0.0598	0.0351
RPL27	0.0612	0.1191
RPL39	0.0616	0.1711
RPLP2	0.0626	0.5767
RPL24	0.0638	0.0875
RPL27A	0.0667	0.0206
RPL35	0.0708	0.1804
RPS15A	0.0748	0.1915
RPS3A	0.0751	0.0895
RPS23	0.0754	0.1384
RPS4X	0.0754	-0.0069
RPL8	0.0761	0.1186
RPS20	0.0777	0.3360
RPL23	0.0784	0.0191
RPL31	0.0798	0.1435
RPL11	0.0802	0.1627
RPL7A	0.0807	0.2953
RPS8	0.0827	0.2151
RPS7	0.0827	0.1056
RPL7	0.0887	0.2130
FAU	0.0900	0.2509
RPL36	0.0918	0.2742
RPL32	0.0932	-0.1833
RPL6	0.0945	-0.0037
RPL18	0.0973	0.2091
RPS11	0.0974	0.1239

Gene symbol	RNA-seq Log2fc	Microarray Log2fc
RPL4	0.0984	0.0754
RPL37	0.0991	-0.0619
RPL5	0.1022	-0.0140
EIF3A	0.1023	-0.0100
RPL34	0.1113	-0.0742
RPS14	0.1132	0.1988
RPS25	0.1169	0.0750
RPS28	0.1173	0.6760
RPL19	0.1185	0.1185
RPL10	0.1187	0.2741
RPS27A	0.1188	0.1977
RPL30	0.1210	0.2704
RPS2	0.1212	0.1375
RPL21	0.1222	0.2936
RPL9	0.1234	0.2967
RPS5	0.1250	0.4106
RPLP0	0.1282	0.1212
RPS3	0.1303	0.0773
EIF3C	0.1303	0.0880
RPL17	0.1318	0.2819
RPS13	0.1325	0.5628
RPL36A	0.1349	0.1592
RPL14	0.1381	0.1100
RPS6	0.1416	0.0458
RPL10A	0.1417	0.2812
RPL22	0.1443	0.5497
RPS27	0.1451	0.4842
RPS21	0.1487	0.1033
RPS29	0.1495	0.2694
RPS17	0.1503	0.3276
EIF4A2	0.1515	0.0020
RPL13	0.1518	0.0424
EIF3J	0.1568	0.1446
EIF3E	0.1603	0.1422
RPL3	0.1608	0.0412
RPS19	0.1655	0.3787
RPL26	0.1674	0.3645
RPS24	0.1687	0.0254
RPL28	0.1731	0.0283
EIF2S2	0.1787	-0.0581
RPS18	0.1798	0.2539
RPL12	0.1841	0.0080
EIF3G	0.1912	0.2367
RPS9	0.2029	0.1276
RPL18A	0.2045	0.3333
RPS16	0.2212	0.1979

Gene symbol	RNA-seq Log2fc	Microarray Log2fc
RPL23AP74	0.2344	NaN
RPL13A	0.2356	0.0872
EIF3F	0.2377	0.2795
RPL10P16	0.2383	NaN
PABPC1	0.2405	-0.0282
RPL10P9	0.2488	NaN
EIF3D	0.2511	0.2320
RPSAP9	0.2729	NaN
RPSAP12	0.2902	NaN
EIF4B	0.3186	-0.1873
RPL21P134	0.3531	NaN
RPS15P5	0.5458	NaN
RPS15AP11	0.5583	NaN
RPL12P2	0.5687	NaN
RPL34P31	0.5802	NaN
RPL26P30	0.8402	NaN
RPL21P119	1.6122	NaN

#### E Kegg Ribosome (Figure 4.6B)

Gene symbol	RNA-seq Log2fc	Microarray Log2fc
RPL10L	-1.7795	0.1917
RPL22L1	-0.0808	-0.0658
RPS26	-0.0691	0.0180
MRPL13	-0.0617	0.1336
RPL41	-0.0534	-0.0004
RPL3L	-0.0528	0.0631
RPL38	-0.0309	0.2651
RPS15	0.0000	0.0384
RPL35A	0.0093	0.1595
RPL26L1	0.0111	-0.0392
RPS27L	0.0116	0.3740
RPL29	0.0232	0.1424
RPL37A	0.0330	0.0947
UBA52	0.0366	0.1356
RPL36AL	0.0470	0.3873
RPL15	0.0484	0.0491
RPLP1	0.0540	0.2898
RPSA	0.0565	0.1233
RPS10	0.0580	0.3682
RPS12	0.0588	0.3266
RPL23A	0.0598	0.0351
RPL27	0.0612	0.1191
RPL39	0.0616	0.1711
RPLP2	0.0626	0.5767

Gene symbol	RNA-seq Log2fc	Microarray Log2fc
RPL24	0.0638	0.0875
RPL27A	0.0667	0.0206
RPL35	0.0708	0.1804
RPS15A	0.0748	0.1915
RPS3A	0.0751	0.0895
RPS23	0.0754	0.1384
RPS4X	0.0754	-0.0069
RPL8	0.0761	0.1186
RPS20	0.0777	0.3360
RPL23	0.0784	0.0191
RPL31	0.0798	0.1435
RPL11	0.0802	0.1627
RPL7A	0.0807	0.2953
RPS8	0.0827	0.2151
RPS7	0.0827	0.1056
RPL7	0.0887	0.2130
FAU	0.0900	0.2509
RPL36	0.0918	0.2742
RPL32	0.0932	-0.1833
RPL6	0.0945	-0.0037
RPL18	0.0973	0.2091
RPS11	0.0974	0.1239
RPL4	0.0984	0.0754
RPL37	0.0991	-0.0619
RPL5	0.1022	-0.0140
RPL34	0.1113	-0.0742
RPS25	0.1169	0.0750
RPS28	0.1173	0.6760
RPL19	0.1185	0.1185
RPL10	0.1187	0.2741
RPS27A	0.1188	0.1977
RPL30	0.1210	0.2704
RPS2	0.1212	0.1375
RPL21	0.1222	0.2936
RPL9	0.1234	0.2967
RPS5	0.1250	0.4106
RPLP0	0.1282	0.1212
RPS3	0.1303	0.0773
RPL17	0.1318	0.2819
RPS13	0.1325	0.5628
RPL36A	0.1349	0.1592
RPL14	0.1381	0.1100
RPS6	0.1416	0.0458
RPL10A	0.1417	0.2812
RPL22	0.1443	0.5497
RPS27	0.1451	0.4842

Gene symbol	RNA-seq Log2fc	Microarray Log2fc
RPS21	0.1487	0.1033
RPS29	0.1495	0.2694
RPS17	0.1503	0.3276
RPL13	0.1518	0.0424
RPL3	0.1608	0.0412
RPS19	0.1655	0.3787
RPL26	0.1674	0.3645
RPS24	0.1687	0.0254
RPL28	0.1731	0.0283
RPS18	0.1798	0.2539
RPL12	0.1841	0.0080
RPS9	0.2029	0.1276
RPL18A	0.2045	0.3333
RPS16	0.2212	0.1979
RPL13A	0.2356	0.0872
RSL24D1	0.2495	0.3539

Gene symbol	RNA-seq Log2fc	Microarray Log2fc
EIF4G1	-0.0183	-0.1292
EIF2B1	-0.0166	-0.1529
EIF5B	-0.0079	-0.0507
EIF2B3	-0.0072	0.0453
RPS15	0.0000	0.0384
EIF2S1	0.0029	-0.0037
EIF3I	0.0039	-0.0487
EIF4A1	0.0041	-0.0019
EIF2S3	0.0054	-0.1099
EIF3B	0.0057	-0.0444
SEC61B	0.0070	0.5524
RPL35A	0.0093	0.1595
RPL26L1	0.0111	-0.0392
EIF3K	0.0160	0.0873
SPCS1	0.0179	0.2092
SRP19	0.0202	0.1428
EIF2B2	0.0207	0.0754
RPL23AP42	0.0209	NaN
RPN2	0.0209	-0.0043
RPL29	0.0232	0.1424
RPL23AP2	0.0304	NaN
RPL37A	0.0330	0.0947
EEF1D	0.0356	0.0747
UBA52	0.0366	0.1356
EIF3FP3	0.0376	NaN
RPN1	0.0378	0.1443
SPCS2	0.0422	0.2838
ETF1	0.0425	0.0694
RPL23AP63	0.0458	NaN
RPL15	0.0484	0.0491
EIF3H	0.0535	0.0636
RPLP1	0.0540	0.2898
RPS27AP11	0.0564	NaN
RPSA	0.0565	0.1233
SEC11A	0.0568	0.1863
RPS10	0.0580	0.3682
RPS12	0.0588	0.3266
RPL23A	0.0598	0.0351
RPL27	0.0612	0.1191
RPL39	0.0616	0.1711
RPLP2	0.0626	0.5767
RPL24	0.0638	0.0875
RPL27A	0.0667	0.0206
EEF1G	0.0706	0.0054
RPL35	0.0708	0.1804
RPS15A	0.0748	0.1915

#### F Reactome Translation (Figure 4.6B)

Gene symbol	RNA-seq Log2fc	Microarray Log2fc
SRP9P1	-1.1494	NaN
FAM153A	-1.0077	NaN
EEF1DP1	-1.0005	NaN
RPS19P3	-1.0003	NaN
RPS29P3	-0.8955	NaN
RPL7AP66	-0.4292	NaN
RPL21P75	-0.3606	NaN
SEC61A2	-0.2742	-0.0853
RPL17P36	-0.2213	NaN
RPL5P1	-0.2151	NaN
EIF4E	-0.1859	-0.1330
RPL23AP18	-0.1821	NaN
SRP54	-0.1330	0.2560
RPS28P7	-0.1103	NaN
RPS26	-0.0691	0.0180
SRP14	-0.0608	0.1440
EIF1AX	-0.0601	-0.0048
RPL41	-0.0534	-0.0004
RPL3L	-0.0528	0.0631
RPS26P28	-0.0505	NaN
RPL21P16	-0.0482	NaN
EIF4H	-0.0409	-0.0577
RPL38	-0.0309	0.2651
SRP9	-0.0246	0.0161
SEC61G	-0.0237	0.8421

Gene symbol	RNA-seq Log2fc	Microarray Log2fc
RPS3A	0.0751	0.0895
RPS23	0.0754	0.1384
RPS4X	0.0754	-0.0069
RPL8	0.0761	0.1186
RPS20	0.0777	0.3360
RPL23	0.0784	0.0191
RPL31	0.0798	0.1435
RPL11	0.0802	0.1627
RPL7A	0.0807	0.2953
SPCS3	0.0812	0.0325
RPS8	0.0827	0.2151
RPS7	0.0827	0.1056
DDOST	0.0845	0.1100
RPL7	0.0887	0.2130
FAU	0.0900	0.2509
RPL36	0.0918	0.2742
RPL32	0.0932	-0.1833
RPL6	0.0945	-0.0037
SRP72	0.0965	0.1484
RPL18	0.0973	0.2091
RPS11	0.0974	0.1239
SEC61A1	0.0979	0.1620
SSR3	0.0982	0.2900
RPL4	0.0984	0.0754
RPL37	0.0991	-0.0619
RPL5	0.1022	-0.0140
EIF3A	0.1023	-0.0100
RPL34	0.1113	-0.0742
EIF2B5	0.1130	0.0838
RPS14	0.1132	0.1988
RPS25	0.1169	0.0750
RPS28	0.1173	0.6760
RPL19	0.1185	0.1185
RPL10	0.1187	0.2741
RPS27A	0.1188	0.1977
RPL30	0.1210	0.2704
RPS2	0.1212	0.1375
RPL21	0.1222	0.2936
RPL9	0.1234	0.2967
RPS5	0.1250	0.4106
SSR2	0.1262	0.2925
RPLP0	0.1282	0.1212
RPS3	0.1303	0.0773
EIF3C	0.1303	0.0880
RPL17	0.1318	0.2819
RPS13	0.1325	0.5628

Gene symbol	RNA-seq Log2fc	Microarray Log2fc
TRAM1	0.1339	0.2692
RPL36A	0.1349	0.1592
RPL14	0.1381	0.1100
RPS6	0.1416	0.0458
RPL10A	0.1417	0.2812
RPL22	0.1443	0.5497
RPS27	0.1451	0.4842
RPS21	0.1487	0.1033
RPS29	0.1495	0.2694
EEF1B2	0.1499	0.0619
RPS17	0.1503	0.3276
EIF4A2	0.1515	0.0020
RPL13	0.1518	0.0424
SRP68	0.1550	0.1639
EIF3J	0.1568	0.1446
EIF3E	0.1603	0.1422
RPL3	0.1608	0.0412
EEF1A1	0.1621	0.0009
RPS19	0.1655	0.3787
RPL26	0.1674	0.3645
RPS24	0.1687	0.0254
RPL28	0.1731	0.0283
SSR4	0.1768	0.3620
EIF2S2	0.1787	-0.0581
RPS18	0.1798	0.2539
EIF5	0.1815	0.3221
RPL12	0.1841	0.0080
EIF3G	0.1912	0.2367
EIF2B4	0.1917	0.0401
RPS9	0.2029	0.1276
RPL18A	0.2045	0.3333
EEF2	0.2053	0.0760
SRPRB	0.2068	0.3781
SSR1	0.2127	0.2241
RPS16	0.2212	0.1979
SEC11C	0.2240	0.6204
RPL23AP74	0.2344	NaN
RPL13A	0.2356	0.0872
EIF3F	0.2377	0.2795
RPL10P16	0.2383	NaN
PABPC1	0.2405	-0.0282
RPL10P9	0.2488	NaN
EIF3D	0.2511	0.2320
RPSAP9	0.2729	NaN
RPSAP12	0.2902	NaN
EIF4B	0.3186	-0.1873



Gene symbol	RNA-seq Log2fc	Microarray Log2fc
RPL21P134	0.3531	NaN
RPS15P5	0.5458	NaN
RPS15AP11	0.5583	NaN
RPL12P2	0.5687	NaN
RPL34P31	0.5802	NaN
EIF4EBP1	0.6551	0.6136
RPL26P30	0.8402	NaN
RPL21P119	1.6122	NaN

**G Hallmark Interferon Alpha Response**  
(Figure 4.7B)

Gene symbol	RNA-seq Log2fc	Microarray Log2fc
SAMD9	-2.7303	-1.2380
IFI44L	-2.6850	-0.3426
IFITM1	-2.5758	0.2133
IFI27	-2.4131	-0.1579
BST2	-1.8617	0.1100
RTP4	-1.7903	0.1747
DDX60	-1.7764	-0.7332
IRF9	-1.6535	-0.3909
IFIT3	-1.6230	-0.3357
ISG15	-1.5818	-0.0356
OASL	-1.5814	-0.3098
CD74	-1.5735	0.0107
IFIT2	-1.4504	-0.6651
CMPK2	-1.4018	-0.1037
SAMD9L	-1.3642	-0.3605
OAS1	-1.2567	-0.2837
RSAD2	-1.2128	-0.1776
TMEM140	-1.1446	-0.0616
GBP4	-1.0748	-0.0660
CXCL10	-1.0077	0.1300
IFI44	-0.9449	-0.7316
IRF7	-0.9153	0.1317
UBA7	-0.8630	0.0937
LPAR6	-0.8589	-0.3565
IL7	-0.7498	-0.0080
PLSCR1	-0.7266	-0.1096
HERC6	-0.6053	-0.4434
SP110	-0.5587	0.0597
IL4R	-0.5292	-0.3141
EIF2AK2	-0.5157	-0.1914
PARP9	-0.4983	-0.4143
IFITM3	-0.4834	0.5178

Gene symbol	RNA-seq Log2fc	Microarray Log2fc
USP18	-0.4721	-0.1432
DHX58	-0.4464	-0.0433
CSF1	-0.3985	-0.0471
PARP12	-0.3927	-0.1028
IFI35	-0.3826	-0.0501
PARP14	-0.3170	-0.0573
LGALS3BP	-0.3161	-0.0374
TDRD7	-0.2983	-0.2165
PSMB9	-0.2862	-0.1327
FAM46A	-0.2842	-0.1882
SELL	-0.2276	-0.3714
CD47	-0.2257	-0.2911
EPSTI1	-0.2179	-0.1821
TRIM21	-0.2110	0.0834
IRF1	-0.2074	0.0380
TRIM5	-0.1820	-0.2579
TRAFD1	-0.1718	-0.2630
MOV10	-0.1693	-0.2120
C1S	-0.1607	-0.3005
STAT2	-0.1572	-0.2582
IFIH1	-0.1410	0.0596
ADAR	-0.1331	-0.0455
IFI30	-0.1231	0.0268
CNP	-0.1153	0.0151
PNPT1	-0.1149	0.1053
OGFR	-0.1065	-0.0170
MX1	-0.1057	-0.0337
PSME2	-0.0883	0.1326
LAP3	-0.0874	0.0697
NCOA7	-0.0690	-0.2672
PSMA3	-0.0597	0.2822
TAP1	-0.0511	0.0361
TRIM25	-0.0372	0.0655
IFITM2	-0.0273	0.2471
B2M	-0.0149	0.4867
ELF1	-0.0141	-0.0458
ISG20	-0.0123	0.0802
RIPK2	-0.0121	-0.1933
PSME1	0.0014	0.1905
CXCL11	0.0284	-0.0132
TRIM26	0.0391	0.0142
IRF2	0.0667	-0.0172
NUB1	0.0674	0.0000
SLC25A28	0.0846	-0.1677
CASP8	0.1239	0.0284
RNF31	0.1251	0.0906

Gene symbol	RNA-seq Log2fc	Microarray Log2fc
NMI	0.1743	0.1277
IL15	0.1927	0.0330
UBE2L6	0.2110	0.2285
LY6E	0.2766	0.3695
WARS	0.3722	0.7438
HLA-C	0.3813	0.3208
TRIM14	0.4097	0.0565
PSMB8	0.4721	0.6219
TXNIP	0.5308	0.1513
GMPR	0.5644	0.2559
BATF2	0.5891	-0.1297
PROCR	0.7654	0.0197
GBP2	0.7844	-0.0856
CCRL2	0.9722	0.3189
LAMP3	1.9013	1.4890

### H Hallmark Estrogen Response Early (Figure 4.7C)

Gene symbol	E2 stim Log2fc	RNA-seq Log2fc	Microarray Log2fc
KLK10	0.1610	-1.5210	-0.0099
CLIC3	-0.1146	-1.5111	-0.1452
DLC1	0.1506	-1.4809	0.0426
TGM2	0.5012	-1.4303	-0.1086
AREG	0.0786	-1.3155	-0.3999
CALB2	0.2397	-0.9960	0.1653
SLC1A1	0.0440	-0.9037	-0.2250
WISP2	0.8711	-0.8951	-0.0707
CALCR	0.8510	-0.8435	-0.5313
FOS	0.4910	-0.7177	-0.7646
KRT13	2.2218	-0.7025	0.0239
EGR3	0.5153	-0.6125	-0.2819
PAPSS2	0.9026	-0.5990	-0.5041
ALDH3B1	-0.1254	-0.5779	-0.1564
STC2	0.9280	-0.5667	-0.5134
ELOVL2	0.3624	-0.5595	-0.8761
FAM134B	0.1387	-0.5411	-0.6177
CHPT1	0.2432	-0.4953	-0.3598
GJA1	-0.1668	-0.4942	-0.0478
SLC26A2	0.1276	-0.4799	-0.1280
ADCY1	0.8436	-0.4599	-0.2417
MYBL1	NaN	-0.4410	-0.5219
FHL2	0.3217	-0.4283	-0.1577
RAPGEFL1	1.7532	-0.4194	-0.5973
TIAM1	0.8982	-0.4057	-0.3269

Gene symbol	E2 stim Log2fc	RNA-seq Log2fc	Microarray Log2fc
RET	1.5570	-0.4028	-0.3223
KCNK15	0.2561	-0.3965	0.0418
INHBB	-0.2229	-0.3926	-0.4360
ARL3	0.5319	-0.3541	-0.1454
ABAT	0.6693	-0.3469	-0.3669
RHOBTB3	-0.0475	-0.3442	-0.3968
KLF10	0.1718	-0.3433	-0.4860
SNX24	0.8031	-0.3221	-0.1110
SLC24A3	0.7964	-0.3152	-0.5471
RARA	0.5141	-0.3068	-0.3971
PDLIM3	0.2173	-0.3010	0.0501
NBL1	0.3198	-0.2980	0.0836
BCL2	0.3098	-0.2898	-0.3738
SYNGR1	-0.1298	-0.2847	-0.1456
CCND1	0.3961	-0.2699	0.0189
CELSR2	0.8620	-0.2628	-0.1864
KCNK5	0.4627	-0.2561	-0.4572
FLNB	0.7179	-0.2523	-0.5145
FKBP5	0.8488	-0.2453	-0.3913
RAB31	1.6962	-0.2324	-0.2000
ADD3	0.1830	-0.2292	-0.5111
HSPB8	1.7115	-0.2292	-0.2181
ENDOD1	0.2124	-0.2221	-0.1924
IL17RB	1.9630	-0.2141	-0.5512
IL6ST	0.2668	-0.2140	-0.1091
ZNF185	0.6865	-0.2105	-0.3830
MICB	1.1499	-0.2104	-0.0877
SFN	-0.3736	-0.2101	-0.3299
NAV2	0.4216	-0.1976	-0.3219
FOXC1	1.4102	-0.1805	-0.0354
NRIP1	0.4248	-0.1793	-0.0335
JAK2	0.3019	-0.1771	-0.1811
HES1	-0.8073	-0.1681	-0.3198
GFRA1	0.5658	-0.1679	-0.1140
CYP26B1	1.2380	-0.1659	0.0998
FDFT1	0.2796	-0.1442	-0.1033
CXCL12	1.4908	-0.1385	-0.8102
PRSS23	0.4748	-0.1360	-0.1042
TGIF2	0.2904	-0.1325	-0.0681
CA12	2.0544	-0.1295	-0.3405
SLC19A2	0.1013	-0.1275	-0.2883
KRT18	-0.3268	-0.1260	-0.1111
RASGRP1	0.5434	-0.1111	-0.1144
TMEM164	0.3725	-0.1053	-0.3650
DYNLT3	0.1184	-0.1046	-0.0594
OLFM1	0.7814	-0.1025	0.0538

Gene symbol	E2 stim Log2fc	RNA-seq Log2fc	Microarray Log2fc
PTGES	1.0002	-0.0965	0.1146
BHLHE40	-0.6084	-0.0955	0.0439
KRT15	0.7001	-0.0952	-0.1490
BCL11B	-0.1093	-0.0934	0.0615
DHRS3	-1.5434	-0.0919	0.0347
SEC14L2	0.6156	-0.0863	0.0159
SLC22A5	0.6826	-0.0811	-0.1810
SLC16A1	-0.0637	-0.0776	-0.1329
MAST4	0.3957	-0.0694	-0.2523
SLC2A1	0.6667	-0.0571	0.0837
SLC9A3R1	1.2688	-0.0567	-0.2644
RBBP8	0.4674	-0.0564	-0.0385
FKBP4	1.3060	-0.0538	-0.0444
HR	-0.1702	-0.0531	-0.1005
SULT2B1	-0.1425	-0.0528	-0.3812
TPBG	0.5997	-0.0464	-0.1740
SIAH2	1.0492	-0.0379	0.0759
KRT8	-0.2343	-0.0249	-0.0548
SEMA3B	0.4916	-0.0183	-0.1365
ELF3	-0.9186	-0.0175	-0.4554
TPD52L1	1.3483	-0.0173	-0.1058
ELF1	0.9302	-0.0141	-0.0458
ITPK1	0.1494	-0.0141	-0.1393
TIPARP	1.2406	-0.0125	0.2159
UNC119	0.6317	-0.0001	0.0032
NPY1R	0.1057	0.0022	-0.4839
DHCR7	0.4877	0.0095	0.0331
INPP5F	-0.1273	0.0100	-0.3132
TFAP2C	0.7622	0.0108	-0.1144
MED24	0.3516	0.0112	-0.0625
PGR	1.5905	0.0120	0.0032
THSD4	-0.0298	0.0123	-0.1570
WFS1	0.3808	0.0134	0.2242
RHOD	0.1414	0.0176	-0.0372
ADCY9	0.0159	0.0216	-0.3469
OLFML3	0.7514	0.0253	0.1509
KLF4	0.7437	0.0263	-0.1140
PEX11A	0.6139	0.0275	-0.0270
TTC39A	0.0803	0.0290	-0.1454
MYOF	0.2563	0.0334	-0.4934
GREB1	1.4805	0.0501	-0.1375
FARP1	0.6152	0.0509	-0.2725
FASN	-0.1628	0.0519	-0.3383
UGCG	0.6413	0.0569	-0.1438
KRT19	0.2621	0.0588	-0.0534
TSKU	0.1025	0.0643	-0.0396

Gene symbol	E2 stim Log2fc	RNA-seq Log2fc	Microarray Log2fc
REEP1	-0.0510	0.0690	-0.0744
IGF1R	0.1312	0.0704	-0.0324
TOB1	-0.1760	0.0738	-0.0833
SLC39A6	0.7533	0.0745	0.0486
TUBB2B	0.8735	0.0758	0.0344
NXT1	0.8003	0.0889	0.1625
OPN3	0.8195	0.0895	-0.1649
OVOL2	0.5312	0.0936	0.0829
PODXL	0.9031	0.0987	0.0317
ESRP2	0.1365	0.1043	-0.1963
ABHD2	0.7359	0.1141	0.0262
NCOR2	-0.0827	0.1219	-0.2367
SCARB1	0.2751	0.1227	-0.0014
PPIF	0.4012	0.1265	-0.0307
ISG20L2	NaN	0.1276	0.0263
IGFBP4	0.4938	0.1356	0.0789
TBC1D30	NaN	0.1385	0.1336
MYBBP1A	0.1590	0.1419	-0.0929
AR	0.1766	0.1434	-0.1838
AFF1	0.1881	0.1483	0.0173
ELOVL5	0.3343	0.1516	0.0994
RRP12	0.5220	0.1525	0.2270
SH3BP5	-0.1181	0.1567	0.2115
AKAP1	0.4458	0.1676	-0.0876
RPS6KA2	0.4636	0.1678	-0.0682
ABLIM1	-0.2810	0.1771	0.1058
KDM4B	0.1068	0.1790	0.0661
SLC37A1	-0.0649	0.1889	0.0627
CLDN7	-0.1970	0.1936	0.2095
CANT1	-0.3134	0.2071	-0.0582
P2RY2	0.1887	0.2157	-0.1494
ASB13	0.4730	0.2221	0.1155
AMFR	0.1513	0.2227	0.1192
SYT12	-0.1197	0.2301	-0.2679
MED13L	0.7885	0.2393	0.0580
ANXA9	0.7573	0.2455	-0.2402
NADSYN1	0.3744	0.2558	-0.0720
FRK	0.3329	0.2592	-0.2250
TMPRSS3	0.2432	0.2794	0.0256
LAD1	0.3154	0.2797	0.0789
SOX3	1.1955	0.3044	0.1672
XBP1	1.0173	0.3094	0.1135
CISH	0.3457	0.3101	0.1234
MLPH	0.0754	0.3109	-0.0819
TFF3	0.4369	0.3157	0.6377
MPPED2	0.0815	0.3362	-0.0716

Gene symbol	E2 stim Log2fc	RNA-seq Log2fc	Microarray Log2fc
SLC27A2	0.6456	0.3365	0.3957
WWC1	1.1112	0.3498	0.0397
FAM102A	0.9054	0.3505	0.2131
KAZN	0.8906	0.3507	NaN
LRIG1	0.6971	0.3586	0.0984
SLC1A4	0.0952	0.3720	0.6137
SLC7A2	0.3003	0.4154	0.8142
SLC7A5	1.3222	0.4168	0.2234
CD44	1.3803	0.4267	-0.1419
MREG	0.8107	0.4392	0.2726
CELSR1	-0.0190	0.4397	0.1894
BLVRB	0.7490	0.4455	0.6431
GAB2	0.4351	0.4544	0.0828
SVIL	0.5430	0.5000	0.2340
MYC	1.3568	0.5187	0.0747
ABCA3	0.5487	0.5398	0.1849
TJP3	NaN	0.5421	0.0931
BAG1	0.1469	0.5574	0.2355
CBFA2T3	-0.1059	0.5812	0.4814
SYBU	1.1075	0.5914	0.2433
MAPT	0.6880	0.6244	-0.0513
SCNN1A	-0.2220	0.6297	0.0708
PDZK1	0.5688	0.6532	0.3958
TFF1	1.6326	0.6745	0.6324
MYB	1.9419	0.6876	0.0387
RAB17	0.5100	0.7064	0.3854
PMAIP1	1.2745	0.7347	0.7747
DHRS2	0.5294	0.9587	1.2313
B4GALT1	-0.0470	0.9615	0.6469
PLA2G16	0.7582	0.9932	0.4943
MUC1	0.1853	1.0470	0.0920
GLA	1.8003	1.2513	0.8888
DEPTOR	1.5016	1.2595	NaN
MSMB	0.4837	1.8795	0.3861
AQP3	0.1594	2.0948	0.4856

**I Hallmark Estrogen Response Late**  
(Figure 4.7C)

Gene symbol	E2 stim Log2fc	RNA-seq Log2fc	Microarray Log2fc
KLK11	1.1232	-2.6484	-0.0786
KLK10	0.1610	-1.5210	-0.0099
CLIC3	-0.1146	-1.5111	-0.1452
TRIM29	0.1210	-1.4917	-0.1388
ACOX2	0.5163	-1.3430	-0.0682

Gene symbol	E2 stim Log2fc	RNA-seq Log2fc	Microarray Log2fc
AREG	0.0786	-1.3155	-0.3999
WISP2	0.8711	-0.8951	-0.0707
ATP2B4	-0.1073	-0.8877	-0.6321
CALCR	0.8510	-0.8435	-0.5313
GJB3	0.1963	-0.7724	-0.1173
UGDH	0.1067	-0.7552	-0.6868
AGR2	0.8505	-0.7435	-0.9379
PCP4	1.5226	-0.7357	0.0180
FOS	0.4910	-0.7177	-0.7646
KRT13	2.2218	-0.7025	0.0239
ST6GALNAC2	0.1315	-0.6832	-0.7858
EGR3	0.5153	-0.6125	-0.2819
PAPSS2	0.9026	-0.5990	-0.5041
ALDH3B1	-0.1254	-0.5779	-0.1564
CHPT1	0.2432	-0.4953	-0.3598
SLC26A2	0.1276	-0.4799	-0.1280
CAV1	0.7625	-0.4565	-0.2195
PTGER3	0.3381	-0.4443	0.2897
PRKAR2B	-0.1084	-0.4256	-0.2197
RAPGEFL1	1.7532	-0.4194	-0.5973
DUSP2	0.2035	-0.4185	-0.2427
TIAM1	0.8982	-0.4057	-0.3269
RET	1.5570	-0.4028	-0.3223
ARL3	0.5319	-0.3541	-0.1454
NAB2	0.6340	-0.3449	-0.2004
KIF20A	-0.0183	-0.3419	-0.5213
GINS2	0.8450	-0.3205	-0.2627
SLC24A3	0.7964	-0.3152	-0.5471
TOP2A	-0.2989	-0.3060	-0.4637
SNX10	0.6014	-0.3036	-0.0331
PDLIM3	0.2173	-0.3010	0.0501
NBL1	0.3198	-0.2980	0.0836
BCL2	0.3098	-0.2898	-0.3738
CDC20	0.2927	-0.2853	-0.2255
CCND1	0.3961	-0.2699	0.0189
CPE	0.3227	-0.2664	-0.2805
TFPI2	0.1865	-0.2633	-0.1836
CELSR2	0.8620	-0.2628	-0.1864
PLK4	0.2977	-0.2616	-0.4285
KCNK5	0.4627	-0.2561	-0.4572
CDC6	0.2857	-0.2558	-0.2611
IDH2	-0.6352	-0.2543	-0.2716
FLNB	0.7179	-0.2523	-0.5145
JAK1	0.5113	-0.2500	-0.3564
FKBP5	0.8488	-0.2453	-0.3913
RAB31	1.6962	-0.2324	-0.2000

Gene symbol	E2 stim Log2fc	RNA-seq Log2fc	Microarray Log2fc
ADD3	0.1830	-0.2292	-0.5111
HSPB8	1.7115	-0.2292	-0.2181
CD9	0.1713	-0.2210	-0.2014
IL17RB	1.9630	-0.2141	-0.5512
IL6ST	0.2668	-0.2140	-0.1091
MICB	1.1499	-0.2104	-0.0877
SFN	-0.3736	-0.2101	-0.3299
CYP4F11	0.1926	-0.2082	0.0356
RNASEH2A	0.3952	-0.2048	-0.2076
EMP2	0.0356	-0.1998	-0.3832
PERP	-0.1007	-0.1978	-0.1492
STIL	0.3832	-0.1954	-0.2810
LAMC2	0.3292	-0.1897	-0.0494
DNAJC12	1.0056	-0.1894	0.2279
FOXC1	1.4102	-0.1805	-0.0354
NRIP1	0.4248	-0.1793	-0.0335
JAK2	0.3019	-0.1771	-0.1811
SCUBE2	-0.0664	-0.1714	-0.4024
LLGL2	0.1827	-0.1692	-0.3219
CYP26B1	1.2380	-0.1659	0.0998
FGFR3	-0.1066	-0.1590	-0.2623
DNAJC1	-0.2681	-0.1558	-0.0321
PDCD4	0.1872	-0.1533	-0.1623
FDFT1	0.2796	-0.1442	-0.1033
MDK	-0.2990	-0.1398	-0.1387
CXCL12	1.4908	-0.1385	-0.8102
PRSS23	0.4748	-0.1360	-0.1042
TSPAN13	0.4521	-0.1327	-0.1374
CA12	2.0544	-0.1295	-0.3405
CACNA2D2	0.0843	-0.1223	0.0783
DYNLT3	0.1184	-0.1046	-0.0594
METTL3	0.0047	-0.1033	-0.4743
OLFM1	0.7814	-0.1025	0.0538
PTGES	1.0002	-0.0965	0.1146
SLC22A5	0.6826	-0.0811	-0.1810
SLC16A1	-0.0637	-0.0776	-0.1329
SLC2A8	0.5237	-0.0700	0.0294
GAL	0.2642	-0.0624	0.0325
SLC9A3R1	1.2688	-0.0567	-0.2644
RBBP8	0.4674	-0.0564	-0.0385
IGSF1	0.3460	-0.0560	-0.2723
CDH1	0.0823	-0.0554	-0.1420
FKBP4	1.3060	-0.0538	-0.0444
HR	-0.1702	-0.0531	-0.1005
SULT2B1	-0.1425	-0.0528	-0.3812
TPBG	0.5997	-0.0464	-0.1740

Gene symbol	E2 stim Log2fc	RNA-seq Log2fc	Microarray Log2fc
XRCC3	0.5973	-0.0454	0.0677
SIAH2	1.0492	-0.0379	0.0759
SGK1	1.7670	-0.0298	-0.0560
HOMER2	NaN	-0.0277	-0.0790
GALE	0.1821	-0.0258	0.0860
SEMA3B	0.4916	-0.0183	-0.1365
TPD52L1	1.3483	-0.0173	-0.1058
ITPK1	0.1494	-0.0141	-0.1393
ISG20	1.3695	-0.0123	0.0802
ZFP36	0.7472	-0.0083	-0.3698
NPY1R	0.1057	0.0022	-0.4839
PTPN6	-0.3088	0.0035	-0.1231
IMPA2	0.4342	0.0083	0.0162
DHCR7	0.4877	0.0095	0.0331
TFAP2C	0.7622	0.0108	-0.1144
PGR	1.5905	0.0120	0.0032
WFS1	0.3808	0.0134	0.2242
SERPINA3	0.4943	0.0136	0.7562
PLXNB1	-0.3860	0.0157	-0.2571
ALDH3A2	0.7637	0.0219	-0.0160
TNNC1	-0.0320	0.0237	0.0191
ETFB	-0.1482	0.0254	0.1864
KLF4	0.7437	0.0263	-0.1140
ASCL1	0.2214	0.0272	0.0018
DLG5	0.1802	0.0333	-0.2064
MYOF	0.2563	0.0334	-0.4934
COX6C	0.0623	0.0483	0.3085
TSTA3	-0.1250	0.0483	0.0818
FARP1	0.6152	0.0509	-0.2725
KRT19	0.2621	0.0588	-0.0534
HPRT1	0.4756	0.0609	0.2330
TOB1	-0.1760	0.0738	-0.0833
MOCS2	-0.1952	0.0778	0.0074
FABP5	0.6991	0.0821	0.6265
NXT1	0.8003	0.0889	0.1625
OPN3	0.8195	0.0895	-0.1649
OVOL2	0.5312	0.0936	0.0829
NMU	0.1606	0.0957	-0.1908
DCXR	0.0536	0.1009	0.0872
LSR	-0.3640	0.1033	0.2752
ABHD2	0.7359	0.1141	0.0262
NCOR2	-0.0827	0.1219	-0.2367
SCARB1	0.2751	0.1227	-0.0014
PPIF	0.4012	0.1265	-0.0307
IGFBP4	0.4938	0.1356	0.0789
BATF	0.2603	0.1462	-0.2531

Gene symbol	E2 stim Log2fc	RNA-seq Log2fc	Microarray Log2fc
AFF1	0.1881	0.1483	0.0173
ELOVL5	0.3343	0.1516	0.0994
PKP3	0.1834	0.1606	-0.0667
MEST	0.1690	0.1627	0.2146
RPS6KA2	0.4636	0.1678	-0.0682
RABEP1	0.3924	0.1692	-0.0417
PLAC1	1.4220	0.1695	-0.2594
SERPINA5	0.4683	0.1754	0.0537
TH	0.9033	0.1914	0.0991
ST14	-0.3649	0.1930	0.0444
MAPK13	0.1398	0.1998	-0.0051
ID2	-0.1846	0.2192	0.5785
AMFR	0.1513	0.2227	0.1192
UNC13B	-0.0188	0.2336	0.1901
SLC29A1	-0.1254	0.2342	-0.0048
ANXA9	0.7573	0.2455	-0.2402
ASS1	0.1160	0.2512	0.1344
CHST8	0.1769	0.2515	0.0427
FRK	0.3329	0.2592	-0.2250
SORD	0.4352	0.2755	0.1809
TMPRSS3	0.2432	0.2794	0.0256
TST	0.4066	0.2916	0.5122
SOX3	1.1955	0.3044	0.1672
XBP1	1.0173	0.3094	0.1135
CISH	0.3457	0.3101	0.1234
TFF3	0.4369	0.3157	0.6377
SLC27A2	0.6456	0.3365	0.3957
CKB	-0.1779	0.3368	0.3085
FAM102A	0.9054	0.3505	0.2131
SLC1A4	0.0952	0.3720	0.6137
PRLR	0.7563	0.3738	0.2900
SLC7A5	1.3222	0.4168	0.2234
CD44	1.3803	0.4267	-0.1419
BLVRB	0.7490	0.4455	0.6431
HSPA4L	0.3460	0.4638	0.5672
ABCA3	0.5487	0.5398	0.1849
TJP3	NaN	0.5421	0.0931
BAG1	0.1469	0.5574	0.2355
CCNA1	0.0476	0.5802	-0.0144
MAPT	0.6880	0.6244	-0.0513
SCNN1A	-0.2220	0.6297	0.0708
PDZK1	0.5688	0.6532	0.3958
TFF1	1.6326	0.6745	0.6324
MYB	1.9419	0.6876	0.0387
SERPINA1	0.1538	0.7824	0.0690
DHRS2	0.5294	0.9587	1.2313

Gene symbol	E2 stim Log2fc	RNA-seq Log2fc	Microarray Log2fc
PLA2G16	0.7582	0.9932	0.4943
LTF	0.0180	1.1393	-0.1308
GLA	1.8003	1.2513	0.8888
HMGCS2	-0.0635	1.3264	-0.0430
S100A9	-0.0005	1.6366	0.3825
CA2	1.0580	1.6530	-0.0088

### J Hallmark Estrogen Response Unique to Early (Figure 4.7D)

Gene symbol	E2 stim Log2fc	RNA-seq Log2fc	Microarray Log2fc
DLC1	0.1506	-1.4809	0.0426
TGM2	0.5012	-1.4303	-0.1086
CALB2	0.2397	-0.9960	0.1653
SLC1A1	0.0440	-0.9037	-0.2250
STC2	0.9280	-0.5667	-0.5134
ELOVL2	0.3624	-0.5595	-0.8761
FAM134B	0.1387	-0.5411	-0.6177
GJA1	-0.1668	-0.4942	-0.0478
ADCY1	0.8436	-0.4599	-0.2417
MYBL1	NaN	-0.4410	-0.5219
FHL2	0.3217	-0.4283	-0.1577
KCNK15	0.2561	-0.3965	0.0418
INHBB	-0.2229	-0.3926	-0.4360
ABAT	0.6693	-0.3469	-0.3669
RHOBTB3	-0.0475	-0.3442	-0.3968
KLF10	0.1718	-0.3433	-0.4860
SNX24	0.8031	-0.3221	-0.1110
RARA	0.5141	-0.3068	-0.3971
SYNGR1	-0.1298	-0.2847	-0.1456
ENDOD1	0.2124	-0.2221	-0.1924
ZNF185	0.6865	-0.2105	-0.3830
NAV2	0.4216	-0.1976	-0.3219
HES1	-0.8073	-0.1681	-0.3198
GFRA1	0.5658	-0.1679	-0.1140
TGIF2	0.2904	-0.1325	-0.0681
SLC19A2	0.1013	-0.1275	-0.2883
KRT18	-0.3268	-0.1260	-0.1111
RASGRP1	0.5434	-0.1111	-0.1144
TMEM164	0.3725	-0.1053	-0.3650
BHLHE40	-0.6084	-0.0955	0.0439
KRT15	0.7001	-0.0952	-0.1490
BCL11B	-0.1093	-0.0934	0.0615
DHRS3	-1.5434	-0.0919	0.0347
SEC14L2	0.6156	-0.0863	0.0159

Gene symbol	E2 stim Log2fc	RNA-seq Log2fc	Microarray Log2fc
MAST4	0.3957	-0.0694	-0.2523
SLC2A1	0.6667	-0.0571	0.0837
KRT8	-0.2343	-0.0249	-0.0548
ELF3	-0.9186	-0.0175	-0.4554
ELF1	0.9302	-0.0141	-0.0458
TIPARP	1.2406	-0.0125	0.2159
UNC119	0.6317	-0.0001	0.0032
INPP5F	-0.1273	0.0100	-0.3132
MED24	0.3516	0.0112	-0.0625
THSD4	-0.0298	0.0123	-0.1570
RHOD	0.1414	0.0176	-0.0372
ADCY9	0.0159	0.0216	-0.3469
OLFML3	0.7514	0.0253	0.1509
PEX11A	0.6139	0.0275	-0.0270
TTC39A	0.0803	0.0290	-0.1454
GREB1	1.4805	0.0501	-0.1375
FASN	-0.1628	0.0519	-0.3383
UGCG	0.6413	0.0569	-0.1438
TSKU	0.1025	0.0643	-0.0396
REEP1	-0.0510	0.0690	-0.0744
IGF1R	0.1312	0.0704	-0.0324
SLC39A6	0.7533	0.0745	0.0486
TUBB2B	0.8735	0.0758	0.0344
PODXL	0.9031	0.0987	0.0317
ESRP2	0.1365	0.1043	-0.1963
ISG20L2	NaN	0.1276	0.0263
TBC1D30	NaN	0.1385	0.1336
MYBBP1A	0.1590	0.1419	-0.0929
AR	0.1766	0.1434	-0.1838
RRP12	0.5220	0.1525	0.2270
SH3BP5	-0.1181	0.1567	0.2115
AKAP1	0.4458	0.1676	-0.0876
ABLIM1	-0.2810	0.1771	0.1058
KDM4B	0.1068	0.1790	0.0661
SLC37A1	-0.0649	0.1889	0.0627
CLDN7	-0.1970	0.1936	0.2095
CANT1	-0.3134	0.2071	-0.0582
P2RY2	0.1887	0.2157	-0.1494
ASB13	0.4730	0.2221	0.1155
SYT12	-0.1197	0.2301	-0.2679
MED13L	0.7885	0.2393	0.0580
NADSYN1	0.3744	0.2558	-0.0720
LAD1	0.3154	0.2797	0.0789
MLPH	0.0754	0.3109	-0.0819
MPPED2	0.0815	0.3362	-0.0716
WWC1	1.1112	0.3498	0.0397

Gene symbol	E2 stim Log2fc	RNA-seq Log2fc	Microarray Log2fc
KAZN	0.8906	0.3507	0.3507
LRIG1	0.6971	0.3586	0.0984
SLC7A2	0.3003	0.4154	0.8142
MREG	0.8107	0.4392	0.2726
CELSR1	-0.0190	0.4397	0.1894
GAB2	0.4351	0.4544	0.0828
SVIL	0.5430	0.5000	0.2340
MYC	1.3568	0.5187	0.0747
CBFA2T3	-0.1059	0.5812	0.4814
SYBU	1.1075	0.5914	0.2433
RAB17	0.5100	0.7064	0.3854
PMAIP1	1.2745	0.7347	0.7747
B4GALT1	-0.0470	0.9615	0.6469
MUC1	0.1853	1.0470	0.0920
DEPTOR	1.5016	1.2595	1.2595
MSMB	0.4837	1.8795	0.3861
AQP3	0.1594	2.0948	0.4856

### K Hallmark Estrogen Response Unique to Late (Figure 4.7D)

Gene symbol	E2 stim Log2fc	RNA-seq Log2fc	Microarray Log2fc
KLK11	1.1232	-2.6484	-0.0786
TRIM29	0.1210	-1.4917	-0.1388
ACOX2	0.5163	-1.3430	-0.0682
ATP2B4	-0.1073	-0.8877	-0.6321
GJB3	0.1963	-0.7724	-0.1173
UGDH	0.1067	-0.7552	-0.6868
AGR2	0.8505	-0.7435	-0.9379
PCP4	1.5226	-0.7357	0.0180
ST6GALNAC2	0.1315	-0.6832	-0.7858
CAV1	0.7625	-0.4565	-0.2195
PTGER3	0.3381	-0.4443	0.2897
PRKAR2B	-0.1084	-0.4256	-0.2197
DUSP2	0.2035	-0.4185	-0.2427
NAB2	0.6340	-0.3449	-0.2004
KIF20A	-0.0183	-0.3419	-0.5213
GINS2	0.8450	-0.3205	-0.2627
TOP2A	-0.2989	-0.3060	-0.4637
SNX10	0.6014	-0.3036	-0.0331
CDC20	0.2927	-0.2853	-0.2255
CPE	0.3227	-0.2664	-0.2805
TFPI2	0.1865	-0.2633	-0.1836
PLK4	0.2977	-0.2616	-0.4285
CDC6	0.2857	-0.2558	-0.2611

Gene symbol	E2 stim Log2fc	RNA-seq Log2fc	Microarray Log2fc
IDH2	-0.6352	-0.2543	-0.2716
JAK1	0.5113	-0.2500	-0.3564
CD9	0.1713	-0.2210	-0.2014
CYP4F11	0.1926	-0.2082	0.0356
RNASEH2A	0.3952	-0.2048	-0.2076
EMP2	0.0356	-0.1998	-0.3832
PERP	-0.1007	-0.1978	-0.1492
STIL	0.3832	-0.1954	-0.2810
LAMC2	0.3292	-0.1897	-0.0494
DNAJC12	1.0056	-0.1894	0.2279
SCUBE2	-0.0664	-0.1714	-0.4024
LLGL2	0.1827	-0.1692	-0.3219
FGFR3	-0.1066	-0.1590	-0.2623
DNAJC1	-0.2681	-0.1558	-0.0321
PDCD4	0.1872	-0.1533	-0.1623
MDK	-0.2990	-0.1398	-0.1387
TSPAN13	0.4521	-0.1327	-0.1374
CACNA2D2	0.0843	-0.1223	0.0783
METTL3	0.0047	-0.1033	-0.4743
SLC2A8	0.5237	-0.0700	0.0294
GAL	0.2642	-0.0624	0.0325
IGSF1	0.3460	-0.0560	-0.2723
CDH1	0.0823	-0.0554	-0.1420
XRCC3	0.5973	-0.0454	0.0677
SGK1	1.7670	-0.0298	-0.0560
HOMER2	NaN	-0.0277	-0.0790
GALE	0.1821	-0.0258	0.0860
ISG20	1.3695	-0.0123	0.0802
ZFP36	0.7472	-0.0083	-0.3698
PTPN6	-0.3088	0.0035	-0.1231
IMPA2	0.4342	0.0083	0.0162
SERPINA3	0.4943	0.0136	0.7562
PLXNB1	-0.3860	0.0157	-0.2571
ALDH3A2	0.7637	0.0219	-0.0160
TNNC1	-0.0320	0.0237	0.0191
ETFB	-0.1482	0.0254	0.1864
ASCL1	0.2214	0.0272	0.0018
DLG5	0.1802	0.0333	-0.2064
COX6C	0.0623	0.0483	0.3085
TSTA3	-0.1250	0.0483	0.0818
HPRT1	0.4756	0.0609	0.2330
MOCS2	-0.1952	0.0778	0.0074
FABP5	0.6991	0.0821	0.6265
NMU	0.1606	0.0957	-0.1908
DCXR	0.0536	0.1009	0.0872
LSR	-0.3640	0.1033	0.2752

Gene symbol	E2 stim Log2fc	RNA-seq Log2fc	Microarray Log2fc
BATF	0.2603	0.1462	-0.2531
PKP3	0.1834	0.1606	-0.0667
MEST	0.1690	0.1627	0.2146
RABEP1	0.3924	0.1692	-0.0417
PLAC1	1.4220	0.1695	-0.2594
SERPINA5	0.4683	0.1754	0.0537
TH	0.9033	0.1914	0.0991
ST14	-0.3649	0.1930	0.0444
MAPK13	0.1398	0.1998	-0.0051
ID2	-0.1846	0.2192	0.5785
UNC13B	-0.0188	0.2336	0.1901
SLC29A1	-0.1254	0.2342	-0.0048
ASS1	0.1160	0.2512	0.1344
CHST8	0.1769	0.2515	0.0427
SORD	0.4352	0.2755	0.1809
TST	0.4066	0.2916	0.5122
CKB	-0.1779	0.3368	0.3085
PRLR	0.7563	0.3738	0.2900
HSPA4L	0.3460	0.4638	0.5672
CCNA1	0.0476	0.5802	-0.0144
SERPINA1	0.1538	0.7824	0.0690
LTF	0.0180	1.1393	-0.1308
HMGCS2	-0.0635	1.3264	-0.0430
S100A9	-0.0005	1.6366	0.3825
CA2	1.0580	1.6530	-0.0088

#### L Charafe Breast Cancer Luminal vs Basal Up (Figure 4.7E)

Gene symbol	RNA-seq Log2fc	Microarray Log2fc
DNALI1	-1.5389	0.2080
DACH1	-1.0176	0.0461
PRRT3	-0.9363	-0.6065
CACNG4	-0.8965	-1.1826
AGR2	-0.7435	-0.9379
PCP4	-0.7357	0.0180
FAM65C	-0.6674	0.0091
CREB3L1	-0.6065	-0.1016
ANKRD30A	-0.6042	-0.0134
CADM1	-0.5931	-0.1872
ANXA6	-0.5892	-0.7257
ELOVL2	-0.5595	-0.8761
RHOH	-0.5449	-0.1971
FAM110B	-0.5182	-0.2723
AFF3	-0.5127	-0.5031



Gene symbol	RNA-seq Log2fc	Microarray Log2fc
C14ORF132	-0.4989	-0.2313
C10ORF82	-0.4850	0.1249
F7	-0.4373	0.0848
INHBB	-0.3926	-0.4360
PRRT2	-0.3636	-0.1446
NBPF1	-0.3373	-0.0215
NUDT4	-0.3232	-0.6339
ETNK2	-0.3219	-0.3386
SLC24A3	-0.3152	-0.5471
MEGF9	-0.3117	-0.6739
CHN2	-0.2954	-0.1813
MB21D2	-0.2870	NaN
DUSP8	-0.2791	-0.0363
PLXNA3	-0.2754	-0.4205
CCND1	-0.2699	0.0189
CXXC5	-0.2690	-0.4316
KLHL22	-0.2420	-0.1485
RHOB	-0.2309	-0.6233
ARID2	-0.2231	-0.1326
TRIL	-0.2168	0.1583
LFNG	-0.2168	-0.3637
DDAH2	-0.2166	-0.3507
POLE	-0.2163	-0.3474
ZNF704	-0.2156	-0.1720
HPN	-0.2062	-0.0339
EMP2	-0.1998	-0.3832
MYO6	-0.1887	-0.4841
INPP5J	-0.1837	-0.4732
SYCP2	-0.1746	-0.5518
FAM46C	-0.1728	-0.1509
SCUBE2	-0.1714	-0.4024
LLGL2	-0.1692	-0.3219
SOX13	-0.1669	-0.2885
DNAJC1	-0.1558	-0.0321
KIF12	-0.1483	0.0108
ZNF24	-0.1457	0.0579
DIP2C	-0.1399	-0.3784
TRIM3	-0.1380	-0.3398
TSPAN13	-0.1327	-0.1374
TGIF2	-0.1325	-0.0681
PLCXD1	-0.1312	-0.2218
STARD10	-0.1301	-0.1989
CA12	-0.1295	-0.3405
CERS2	-0.1277	NaN
TGFB3	-0.1251	-0.4983
ERBB3	-0.1242	-0.3512

Gene symbol	RNA-seq Log2fc	Microarray Log2fc
MAPK9	-0.1229	-0.2277
CACNA2D2	-0.1223	0.0783
GRAMD4	-0.1196	-0.1415
KIAA0232	-0.1193	-0.2341
EPS8L1	-0.1176	-0.1799
FRS2	-0.1170	-0.0368
MAGED2	-0.1147	-0.0715
MGAT4A	-0.1064	-0.1929
GPRC5C	-0.1060	-0.0593
FGFR4	-0.1046	-0.0513
CFD	-0.1044	-0.1624
CACNA1D	-0.1036	-0.3021
EVL	-0.1022	-0.1969
ZNF398	-0.1010	0.0011
CTXN1	-0.0991	0.0010
KCTD15	-0.0916	0.0173
POGZ	-0.0871	-0.2008
EFR3B	-0.0839	-0.1054
TTC3	-0.0826	-0.1533
LMCD1	-0.0825	0.0695
CHTOP	-0.0808	NaN
HNRNPA2B1	-0.0762	-0.2011
FTX	-0.0745	NaN
API5	-0.0707	0.0422
IQSEC1	-0.0706	-0.2748
DAAM1	-0.0677	-0.1902
NKAIN1	-0.0592	-0.0221
SLC9A3R1	-0.0567	-0.2644
FKBP4	-0.0538	-0.0444
SFI1	-0.0529	-0.2923
REEP5	-0.0520	-0.1001
ABHD12	-0.0505	-0.0879
FOXA1	-0.0504	-0.1685
ZFYVE16	-0.0503	-0.2205
DSCAM-AS1	-0.0494	NaN
SLC25A44	-0.0486	-0.0548
KAT6B	-0.0416	NaN
C9ORF152	-0.0373	-0.6468
FZD4	-0.0345	0.0005
SMARCC2	-0.0338	-0.2481
RHBDF1	-0.0337	-0.2417
ZNF84	-0.0311	0.0513
IQCE	-0.0304	-0.1543
UAP1L1	-0.0301	-0.0570
GART	-0.0300	-0.0080
ZMIZ1	-0.0297	-0.2846

Gene symbol	RNA-seq Log2fc	Microarray Log2fc
BPTF	-0.0296	-0.0790
GNA12	-0.0264	-0.0604
CERS6	-0.0250	NaN
ENPP1	-0.0240	-0.1625
ANKRD13D	-0.0238	-0.1284
CSRNP2	-0.0216	0.0167
ABHD11	-0.0202	-0.4011
FUS	-0.0164	-0.3172
TAPT1	-0.0145	-0.0449
DHRS13	-0.0145	0.0356
TMEM150C	-0.0141	NaN
TADA2B	-0.0132	0.0628
ISG20	-0.0123	0.0802
MCCC2	-0.0117	-0.0285
NDUFS8	-0.0107	0.1200
VPS72	-0.0107	0.0907
GTF3C1	-0.0096	-0.0807
ARF3	-0.0078	-0.0413
USP7	-0.0069	-0.0756
ESR1	-0.0067	-0.0701
TESK1	-0.0057	-0.0165
ZBTB42	-0.0043	-0.4111
RDH13	-0.0034	0.0303
GPD1L	-0.0011	-0.2330
CACYBP	-0.0007	0.0101
ZNF12	-0.0001	0.1391
SNED1	0.0000	-0.1176
CEP350	0.0029	0.0543
PPP2R2C	0.0039	-0.1286
C17ORF62	0.0051	-0.0510
EIF3B	0.0057	-0.0444
SLC16A6	0.0059	-0.2578
SH3GLB2	0.0064	-0.2317
LRP3	0.0074	0.1346
TNRC18	0.0077	-0.0354
GSPT1	0.0087	-0.1016
C7ORF26	0.0088	-0.1006
ASH1L	0.0092	0.0063
ERGIC1	0.0096	-0.1767
ULK1	0.0097	0.0012
SNX27	0.0109	-0.0751
TRAPPC9	0.0112	-0.1464
SLC25A29	0.0114	-0.1494
PGR	0.0120	0.0032
THSD4	0.0123	-0.1570
TMEM80	0.0124	0.0313

Gene symbol	RNA-seq Log2fc	Microarray Log2fc
TBL1X	0.0131	-0.0528
MDM4	0.0132	-0.1325
WFS1	0.0134	0.2242
CRNKL1	0.0158	-0.0187
THUMPD1	0.0177	0.1235
GATA3	0.0181	-0.2075
ATP6AP1	0.0203	0.0574
ATXN7L3B	0.0208	0.0883
DLG3	0.0209	-0.1798
KLF2	0.0215	-0.0140
KLHDC9	0.0222	0.1957
USP3	0.0256	0.0416
POMT1	0.0256	0.0999
MYO5B	0.0257	-0.1037
DENND1A	0.0262	-0.0695
PRR14	0.0275	-0.0506
TTC39A	0.0290	-0.1454
PRRC2C	0.0293	NaN
TBX3	0.0305	0.0079
DDX42	0.0318	-0.1940
GGA3	0.0331	-0.0762
TMBIM6	0.0388	0.0679
LUC7L3	0.0392	-0.3661
ZNF444	0.0399	0.0434
CDC42SE1	0.0400	-0.0580
PGGT1B	0.0405	-0.0073
SECISBP2	0.0437	-0.1005
LONRF2	0.0469	0.0799
SLC26A11	0.0486	0.1479
CACNB3	0.0493	-0.3349
PLA2G12A	0.0515	0.1273
DENND4B	0.0524	-0.1194
RUSC1	0.0525	-0.0247
UBN1	0.0528	0.0615
TMEM57	0.0532	0.1606
EPN3	0.0547	-0.0792
USP42	0.0573	0.0369
KRT19	0.0588	-0.0534
RBAK	0.0594	0.1845
RNF103	0.0606	0.0200
GARNL3	0.0608	-0.1516
ACVR1B	0.0621	0.1556
TRPS1	0.0621	-0.1769
SLC38A1	0.0636	-0.0134
KIAA1211	0.0656	0.1283
MGRN1	0.0660	-0.0054

Gene symbol	RNA-seq Log2fc	Microarray Log2fc
RALGAPA1	0.0673	0.2604
ADCY6	0.0682	-0.2997
RGL2	0.0688	-0.1662
PATZ1	0.0695	-0.0910
P4HTM	0.0695	0.0436
ASB8	0.0706	0.0698
CSNK1D	0.0715	-0.0516
NACA	0.0737	0.0399
TOB1	0.0738	-0.0833
PTPRF	0.0744	-0.0161
STRBP	0.0763	0.0762
BAZ2A	0.0791	0.0569
MYEF2	0.0814	-0.0987
ZNF467	0.0827	-0.1821
SCYL3	0.0827	0.0051
KIAA0040	0.0832	0.1352
SERF2	0.0834	0.0810
ARFIP2	0.0847	0.1233
ARHGEF26	0.0847	0.0665
CCDC117	0.0848	0.1589
PPP1R16A	0.0903	0.0043
PCBP2	0.0917	-0.0900
ONECUT2	0.0922	0.1212
SBK1	0.0929	0.1556
SHANK2	0.0932	-0.2212
LARP4B	0.0942	-0.0033
RABEP2	0.0945	-0.1078
RAB11FIP3	0.0967	0.0166
IRGQ	0.0996	0.0009
ZNF74	0.1038	-0.0183
ESRP2	0.1043	-0.1963
TNIP1	0.1055	0.0104
DNAJA4	0.1069	0.1407
BCOR	0.1071	0.0534
SDCCAG3	0.1107	0.1324
TLE3	0.1157	-0.0045
RHPN1	0.1160	-0.1020
VPS37C	0.1164	0.1002
HPX	0.1167	-0.0414
MIF4GD	0.1170	0.2065
ZNF703	0.1215	0.0015
KIAA0556	0.1268	0.0919
CYB561	0.1279	0.1716
PREX1	0.1306	-0.1310
SRRM2	0.1310	0.0093
PBX1	0.1312	-0.0126

Gene symbol	RNA-seq Log2fc	Microarray Log2fc
STRADA	0.1330	0.1055
TBC1D30	0.1385	0.1336
CIRBP	0.1396	0.0816
SIDT2	0.1397	-0.1429
AR	0.1434	-0.1838
SYNGR2	0.1446	0.1794
PLEKHH1	0.1450	-0.1556
HIP1R	0.1508	0.1705
GGA1	0.1563	0.0810
HMG20B	0.1565	0.0065
NLK	0.1570	0.1593
CYHR1	0.1581	0.1730
HEXDC	0.1597	-0.0362
SLC35A1	0.1600	0.1518
SEC16A	0.1642	0.1756
AVL9	0.1644	0.1522
NUCB2	0.1667	0.2606
MARS	0.1686	0.2292
CTNND2	0.1691	-0.2568
VIPR1	0.1710	0.0334
NME3	0.1769	0.0031
RAB3D	0.1770	-0.0498
KDM4B	0.1790	0.0661
ZNF296	0.1817	0.0129
LZTR1	0.1820	-0.0908
SLC38A10	0.1870	0.1142
RSAD1	0.1888	0.0400
SLC37A1	0.1889	0.0627
GAMT	0.1891	0.0339
CAMSAP3	0.1898	NaN
TBC1D16	0.1933	0.1246
SPATA2L	0.1996	0.0332
ARRB1	0.1996	-0.0226
INTS3	0.2031	0.0333
CANT1	0.2071	-0.0582
GALNT6	0.2098	-0.0001
CREB3L4	0.2106	-0.1554
TMEM229B	0.2314	0.0714
ATP6V0E2	0.2366	-0.0085
PI4KA	0.2377	0.0154
KIFC2	0.2414	-0.2040
ANXA9	0.2455	-0.2402
PYCR1	0.2490	0.3979
ASTN2	0.2531	0.1841
SOX12	0.2549	-0.0239
GARS	0.2553	0.3188

Gene symbol	RNA-seq Log2fc	Microarray Log2fc
RAB40C	0.2605	0.2747
GPR160	0.2607	0.0293
CLSTN2	0.2672	-0.0408
SLC2A10	0.2714	-0.0350
TMEM184A	0.2800	0.0623
KLRG2	0.2800	0.1898
FBRSL1	0.2861	0.0625
C17ORF58	0.2865	-0.0676
TTC9	0.2909	-0.0755
ZDHHC8P1	0.2944	-0.0020
RALGPS1	0.3007	-0.0761
IVD	0.3024	0.1259
XBP1	0.3094	0.1135
CISH	0.3101	0.1234
MLPH	0.3109	-0.0819
TFF3	0.3157	0.6377
SFMBT2	0.3248	0.1850
PCK2	0.3452	0.6460
MYCN	0.3491	0.0461
FRMD4A	0.3575	-0.0770
SLC1A4	0.3720	0.6137
SIDT1	0.3722	0.2548
PRLR	0.3738	0.2900
SPDEF	0.3772	0.2238
HK2	0.3867	0.2005
RIIAD1	0.3928	NaN
KIAA1324	0.4295	0.0600
ATP8B1	0.4342	-0.1368
SPTLC2	0.4353	0.4012
ICA1	0.4377	0.6008
GOLT1A	0.4437	0.4374
TC2N	0.4753	0.2621
C4ORF19	0.4777	0.3345
LNX1	0.5181	0.3056
ABCA3	0.5398	0.1849
TJP3	0.5421	0.0931
DOPEY2	0.5480	0.2993
ABCG1	0.5508	0.5841
NPDC1	0.5688	0.3062
MXRA8	0.5802	0.0806
MAPT	0.6244	-0.0513
TFF1	0.6745	0.6324
TSPAN15	0.6761	0.3538
MYB	0.6876	0.0387
RAB17	0.7064	0.3854
CAPN9	0.7193	0.0145

Gene symbol	RNA-seq Log2fc	Microarray Log2fc
ATP2C2	0.7782	0.6731
SLC4A8	0.7833	0.2422
SLC7A8	0.9103	0.8523
RSPH1	1.1052	0.1256
DEGS2	1.1826	-0.1053
DEPTOR	1.2595	NaN
RND1	1.3252	1.4633
HMGCS2	1.3264	-0.0430
BCAS1	1.3854	0.6112
SLC44A4	1.7669	0.0177
BLNK	2.2669	0.8261

#### M LTED Down (Figure 4.8C)

Gene symbol	RNA-seq Log2fc	Microarray Log2fc
SAMD9	-2.7303	-1.2380
IFI44L	-2.6850	-0.3426
IFIT1	-2.6276	-1.0791
IFITM1	-2.5758	0.2133
OAS2	-2.4286	-0.3592
IFI27	-2.4131	-0.1579
BST2	-1.8617	0.1100
DDX60	-1.7764	-0.7332
IFI6	-1.7008	0.1426
IRF9	-1.6535	-0.3909
IFIT3	-1.6230	-0.3357
ISG15	-1.5818	-0.0356
OASL	-1.5814	-0.3098
DLC1	-1.4809	0.0426
DDX58	-1.3537	-0.5989
OAS1	-1.2567	-0.2837
GEM	-1.2450	-0.1954
HERC5	-1.1799	-0.4027
CALB2	-0.9960	0.1653
IFI44	-0.9449	-0.7316
IRF7	-0.9153	0.1317
SLC1A1	-0.9037	-0.2250
WISP2	-0.8951	-0.0707
LGALS1	-0.8794	-0.2195
CALCR	-0.8435	-0.5313
NRP1	-0.7963	-0.9372
CRABP1	-0.7884	0.0010
OAS3	-0.7591	-0.3878
PLSCR1	-0.7266	-0.1096
EFEMP1	-0.6964	-0.9653

Gene symbol	RNA-seq Log2fc	Microarray Log2fc
RPS6KA3	-0.6769	-0.5949
EGR3	-0.6125	-0.2819
HERC6	-0.6053	-0.4434
MATN2	-0.5671	-0.7351
STC2	-0.5667	-0.5134
ELOVL2	-0.5595	-0.8761
SP110	-0.5587	0.0597
EPB41L2	-0.5508	-0.8140
C16ORF45	-0.5404	0.2222
EIF2AK2	-0.5157	-0.1914
USP18	-0.4721	-0.1432
CAV1	-0.4565	-0.2195
HIST1H2AC	-0.4537	-0.3797
FHL2	-0.4283	-0.1577
PRKAR2B	-0.4256	-0.2197
DTNA	-0.4216	-0.5806
DUSP2	-0.4185	-0.2427
ELF4	-0.3936	-0.4715
PARP12	-0.3927	-0.1028
CNN2	-0.3853	-0.5761
IFI35	-0.3826	-0.0501
HIST1H3H	-0.3565	-0.0107
ANXA3	-0.3499	-0.6154
RHOBTB3	-0.3442	-0.3968
JAG1	-0.3345	-0.2384
IFIT5	-0.3315	-0.2084
LGALS3BP	-0.3161	-0.0374
PHF11	-0.3153	-0.0216
SLC24A3	-0.3152	-0.5471
PDLIM3	-0.3010	0.0501
TDRD7	-0.2983	-0.2165
BCL2	-0.2898	-0.3738
PSMB9	-0.2862	-0.1327
SLC6A14	-0.2843	-1.0337
WIPF1	-0.2694	-0.1346
TFPI2	-0.2633	-0.1836
SYTL2	-0.2563	-0.6722
SP100	-0.2407	0.1066
RAB31	-0.2324	-0.2000
ADD3	-0.2292	-0.5111
HIST1H4H	-0.2234	-0.0271
SAMHD1	-0.2205	-0.0001
HIST1H2AG	-0.2172	-0.4961
PTPRK	-0.2141	-0.2263
PTBP2	-0.2112	-0.1978
MICB	-0.2104	-0.0877

Gene symbol	RNA-seq Log2fc	Microarray Log2fc
HIST1H2BG	-0.2064	-0.1738
NAV2	-0.1976	-0.3219
ARHGAP12	-0.1921	-0.2521
CAP2	-0.1911	-0.1967
RFWD3	-0.1896	-0.1263
TFAP2A	-0.1882	-0.0921
ID1	-0.1846	0.0715
NRIP1	-0.1793	-0.0335
JAK2	-0.1771	-0.1811
HES1	-0.1681	-0.3198
GFRA1	-0.1679	-0.1140
DSCAM	-0.1501	-0.4277
LTN1	-0.1436	-0.1515
IFIH1	-0.1410	0.0596
MDK	-0.1398	-0.1387
NCBP1	-0.1394	-0.1553
PRSS23	-0.1360	-0.1042
ZNF107	-0.1325	-0.1551
CA12	-0.1295	-0.3405
LPGAT1	-0.1295	-0.0837
MYO10	-0.1294	0.1134
LAMA3	-0.1266	-0.3014
CSTF2T	-0.1220	-0.1524
ZNF43	-0.1033	0.0350
PPM1E	-0.0970	0.1728
PTGES	-0.0965	0.1146
PCSK6	-0.0933	-0.1233
SLC16A7	-0.0862	-0.3786
FTO	-0.0809	-0.3344
SLC16A1	-0.0776	-0.1329
RIF1	-0.0657	-0.0708
GAL	-0.0624	0.0325
ZNF273	-0.0587	0.1772
IGSF1	-0.0560	-0.2723
LGR4	-0.0524	-0.0931
TAP1	-0.0511	0.0361
CBLL1	-0.0405	0.1509
GTF2I	-0.0400	-0.2107
BCL2L11	-0.0393	-0.2797
ZNF84	-0.0311	0.0513
NIP7	-0.0307	-0.0261
GART	-0.0300	-0.0080
GRB10	-0.0297	0.3443
PSPC1	-0.0277	-0.1318
DNAJB14	-0.0273	-0.0331
COL9A3	-0.0268	0.0737

Gene symbol	RNA-seq Log2fc	Microarray Log2fc
BHLHE41	-0.0259	-0.2260
NUP160	-0.0209	0.0709
PTER	-0.0188	-0.0727
C18ORF8	-0.0095	0.0116
JMJD1C	-0.0065	-0.0805
NPY1R	0.0022	-0.4839
SATB2	0.0035	-0.1485
MPHOSPH8	0.0061	-0.1526
INPP5F	0.0100	-0.3132
CHD1	0.0132	-0.0363
POT1	0.0241	0.0158
KLF4	0.0263	-0.1140
ASCL1	0.0272	0.0018
USP25	0.0272	-0.0851
ALG13	0.0349	0.1146
MTMR6	0.0417	0.0353
XK	0.0475	0.0172
GREB1	0.0501	-0.1375
ARMCX5	0.0532	0.0291
NFIL3	0.0560	0.1758
CHN1	0.0560	-0.0254
UGCG	0.0569	-0.1438
PEG10	0.0625	-0.1236
IGF1R	0.0704	-0.0324
POLI	0.0710	-0.0753
RNF6	0.0783	0.0711
N4BP2L2	0.0783	0.1524
TAF12	0.0871	0.0684
ZMYM2	0.0886	-0.0577
ZNF268	0.0922	0.2086
ZNF140	0.1038	0.5053
UTP14A	0.1046	0.1344
RRN3	0.1057	0.3906
NOTCH1	0.1089	-0.0256
FKTN	0.1125	0.0360
CDK8	0.1198	0.1871
NCOR2	0.1219	-0.2367
CASP8	0.1239	0.0284
CEBPG	0.1277	0.2617
TSPO	0.1312	0.0474
HSPA13	0.1320	0.4587
ATP8A2	0.1416	-0.1302
PLCB1	0.1434	-0.0824
AMMECR1	0.1567	0.0726
GADD45A	0.1668	0.8621
NMI	0.1743	0.1277

Gene symbol	RNA-seq Log2fc	Microarray Log2fc
IFRD1	0.1780	0.2959
CSTA	0.1841	-0.0812
MOCOS	0.1961	0.3285
UBE2L6	0.2110	0.2285
GCA	0.2664	0.3780
SOX3	0.3044	0.1672
KAZN	0.3507	NaN
ATF3	0.3716	-0.0851
TRIM14	0.4097	0.0565
TM4SF1	0.4292	0.0272
AMIGO2	0.4770	0.1366
SVIL	0.5000	0.2340
CBFA2T3	0.5812	0.4814
MGP	0.6221	0.7903
PDZK1	0.6532	0.3958
IL24	0.6805	0.1451
PMAIP1	0.7347	0.7747
SLC7A11	0.8774	1.3100
PSAT1	1.1827	0.3313
LAMP3	1.9013	1.4890

#### N LTED Up (Figure 4.8C)

Gene symbol	RNA-seq Log2fc	Microarray Log2fc
CLIC3	-1.5111	-0.1452
ACOX2	-1.3430	-0.0682
DAB2	-0.9809	-0.1745
ACTG2	-0.8807	-0.2081
AOX1	-0.7992	-0.3799
PCP4	-0.7357	0.0180
MAOA	-0.7325	0.0194
BMP7	-0.6770	-0.3302
RLN2	-0.6510	0.0002
PCDH9	-0.6212	-0.0484
SEPT10	-0.5945	0.1850
ANXA6	-0.5892	-0.7257
MMP9	-0.5856	-0.0464
PTGER4	-0.4809	-0.0833
PPP1R3C	-0.4804	-0.9157
SLC26A2	-0.4799	-0.1280
EDN1	-0.4551	-0.8572
TNS3	-0.4528	-0.6827
SASH1	-0.3901	-0.1395
TNFRSF11B	-0.3837	-0.2016
DPYSL2	-0.3798	-0.2838

Gene symbol	RNA-seq Log2fc	Microarray Log2fc
PBK	-0.3711	-0.4766
SORL1	-0.3353	-0.2749
RIN2	-0.3271	-0.3298
CLMN	-0.3259	-0.7033
NUDT4	-0.3232	-0.6339
BTG1	-0.3054	-0.1625
ABCG2	-0.2761	-0.4736
TK1	-0.2694	-0.2173
FERMT2	-0.2607	-0.4128
EHD3	-0.2397	-0.2601
GGH	-0.2127	0.0592
CEP76	-0.2122	0.0426
SLC12A2	-0.2117	-0.3679
TNNT1	-0.2073	-0.0990
CYP1B1	-0.1774	-0.2491
ECI2	-0.1748	NaN
FAM46C	-0.1728	-0.1509
MYO5A	-0.1597	-0.2496
SGPP1	-0.1552	-0.1058
ACOX1	-0.1540	-0.2311
SLC9A6	-0.1507	0.0424
DIAPH2	-0.1477	-0.0928
CDKN3	-0.1432	-0.1865
SEC23A	-0.1430	-0.1278
APPL2	-0.1383	-0.4697
RAB27A	-0.1323	-0.0736
POLE2	-0.1320	-0.2730
CHST11	-0.1259	-0.5779
DEGS1	-0.1216	-0.0985
EPS8L1	-0.1176	-0.1799
MGAT4A	-0.1064	-0.1929
KIN	-0.1063	0.2017
FGFR4	-0.1046	-0.0513
INSIG2	-0.0998	0.0598
B4GALT5	-0.0987	-0.2398
SECISBP2L	-0.0973	-0.0693
HMGCS1	-0.0853	0.0641
DSC2	-0.0841	0.2842
MRPL13	-0.0617	0.1336
STAU2	-0.0533	-0.0979
TMC6	-0.0502	-0.0826
PLAGL1	-0.0492	-0.2436
ASRGL1	-0.0457	-0.0568
ULBP2	-0.0444	0.1524
BACE2	-0.0442	0.0433
PHACTR2	-0.0437	-0.0516

Gene symbol	RNA-seq Log2fc	Microarray Log2fc
RMND1	-0.0221	-0.1963
MICAL3	-0.0201	0.0952
MEIS2	-0.0196	-0.3140
HRAS	-0.0168	0.0606
NUP43	-0.0140	-0.1360
MARCKS	-0.0137	0.0947
ORAI3	-0.0025	-0.0075
LMO4	-0.0024	0.1784
P4HA2	-0.0010	0.2155
RHOQ	0.0045	0.1251
NPC2	0.0046	0.5589
EPB41L1	0.0079	-0.5455
RAB4A	0.0093	-0.0019
EPHX1	0.0134	-0.0281
CSGALNACT1	0.0167	-0.1343
SOAT1	0.0201	-0.1427
PJA2	0.0207	0.1329
GSTZ1	0.0219	0.0220
DRAM1	0.0268	-0.2850
SCARA3	0.0275	0.0461
AKAP12	0.0321	0.0531
CD302	0.0325	0.1444
GBAS	0.0338	-0.0453
PGM1	0.0340	-0.0268
GLTP	0.0367	-0.1237
GPX3	0.0383	0.1054
P2RX4	0.0396	-0.0268
UHRF1BP1L	0.0412	-0.0881
UBE3B	0.0413	0.0431
KIF13B	0.0461	-0.0131
FBXL18	0.0486	-0.0239
GTDC1	0.0530	-0.0496
EPN3	0.0547	-0.0792
SERHL2	0.0557	-0.1438
UST	0.0568	-0.1130
REEP1	0.0690	-0.0744
UROS	0.0718	0.0109
LRP8	0.0773	-0.1565
TIMM13	0.0785	0.2994
ATP6V1H	0.0808	0.1822
GOLM1	0.0842	-0.1744
PSTPIP2	0.0846	0.0924
GCNT1	0.0861	-0.0995
QPRT	0.0872	0.0140
OPN3	0.0895	-0.1649
EFCAB11	0.0917	NaN

Gene symbol	RNA-seq Log2fc	Microarray Log2fc
ALDH4A1	0.0929	0.0553
KLHDC2	0.0987	0.1052
PRPS1	0.1034	-0.1467
CGRRF1	0.1119	0.2743
DHRS7	0.1153	0.0352
PTOV1	0.1171	0.1930
ENOSF1	0.1180	-0.1609
C1ORF115	0.1206	-0.0108
PDE8A	0.1246	-0.0360
PROS1	0.1271	0.1977
TFPT	0.1284	0.1253
RCAN1	0.1383	0.0737
KATNA1	0.1587	0.0872
NUMB	0.1607	0.1818
NUCB2	0.1667	0.2606
LONP2	0.1689	-0.0117
ASPH	0.1691	0.0611
PLCB4	0.1723	-0.2779
MAP3K5	0.1813	0.1147
PPFIA3	0.1860	-0.2377
ABCC3	0.1907	0.0639
TBC1D8	0.2073	-0.0013
PRKCH	0.2081	0.0868
CREB3L2	0.2259	0.4702
GALNT12	0.2276	-0.0891
UNC13B	0.2336	0.1901
ANKMY2	0.2441	0.2944
BNIP3L	0.2469	0.3424
SLC2A10	0.2714	-0.0350
SORD	0.2755	0.1809
MBP	0.2763	0.0757
ADORA1	0.2887	0.3453
IL1R1	0.2901	-0.3124
TST	0.2916	0.5122
PNRC1	0.3006	0.2150
ATP6V0A4	0.3010	-0.6855
EPAS1	0.3025	-0.0251
PIGH	0.3117	0.1379
GCLM	0.3285	0.5886
RAB40B	0.3533	0.0077
CXCR4	0.4320	0.0674
SPTLC2	0.4353	0.4012
DUSP4	0.4631	0.3549
SYT17	0.4747	0.0859
C4ORF19	0.4777	0.3345
FAM117A	0.5301	0.2732

Gene symbol	RNA-seq Log2fc	Microarray Log2fc
NOVA1	0.5632	0.0282
RRAGD	0.6039	0.4922
CALML5	0.6563	0.0587
CYFIP2	0.6974	0.6001
GNA14	0.7163	0.3043
SDCBP	0.7227	1.4960
FA2H	0.7457	0.0209
SERHL	0.7747	-0.0106
ALDH3B2	0.7811	0.5256
SERPINA1	0.7824	0.0690
RPS6KA5	0.8222	0.7078
RFTN1	0.8308	-0.0184
CEACAM6	0.8664	-0.0999
EHF	0.9806	-0.3399
HMGCS2	1.3264	-0.0430
IGFBP5	1.4985	0.7485
HSD17B14	2.0346	0.0424
AQP3	2.0948	0.4856
ELF5	3.1058	1.9343
PIP	3.2859	0.1412
KMO	3.2994	0.7223
VTCN1	3.8416	2.0284

**O MYC Overexpression Down  
(Figure 4.8D)**

Gene symbol	RNA-seq Log2fc	Microarray Log2fc
SAMD9	-2.7303	-1.2380
CD1D	-1.8766	0.0807
EREG	-1.7187	-0.1399
TGM2	-1.4303	-0.1086
C11ORF72	-1.2544	-0.0978
PTPRE	-1.0532	-0.3629
REC8	-1.0431	-0.0263
SDR16C5	-1.0077	-0.1389
CARD16	-1.0077	0.0891
ZNF334	-0.9912	0.0313
HEG1	-0.9736	-0.9410
TLR2	-0.9008	-0.2831
HOXC4	-0.8797	NaN
SLITRK6	-0.8738	-1.0165
MTMR11	-0.8150	-0.3199
NRP1	-0.7963	-0.9372
WNT5B	-0.7190	-0.1250
PBX4	-0.6701	-0.0383



Gene symbol	RNA-seq Log2fc	Microarray Log2fc
PLEKHG2	-0.6605	-0.4246
KRT80	-0.5982	-0.2573
MMP9	-0.5856	-0.0464
MCHR1	-0.4850	-0.1191
TGM5	-0.4850	0.0810
PLAT	-0.4790	-0.1612
TNFAIP2	-0.4759	-0.1409
HIST1H2AC	-0.4537	-0.3797
PPARD	-0.4391	-0.1563
KLF12	-0.4092	-0.0929
SYT1	-0.3848	-0.2920
MEGF6	-0.3680	-0.0802
PRRT2	-0.3636	-0.1446
AGER	-0.3613	-0.1779
GINS1	-0.3366	-0.4159
ANGPTL4	-0.3290	0.1026
KATNAL1	-0.3283	-0.3773
NGFR	-0.3216	0.0048
MICAL1	-0.3051	-0.2613
PSMC3IP	-0.3027	-0.2484
ACAT2	-0.3004	-0.6219
CCDC82	-0.2862	-0.1459
SLC6A14	-0.2843	-1.0337
LYPD3	-0.2734	-0.2721
WDR47	-0.2610	-0.1510
PLEKHA2	-0.2451	-0.3165
IL6	-0.2285	-0.0647
CDCA8	-0.2256	-0.2681
HIP1	-0.2130	-0.3030
RAP1A	-0.2127	-0.0097
C10ORF10	-0.2126	-0.0015
PAK3	-0.2073	0.0284
ASAP3	-0.1908	-0.2560
RNF19A	-0.1903	0.0283
ABHD3	-0.1781	0.0815
ARRDC3	-0.1675	-0.1345
PBX3	-0.1619	-0.2605
TUBA1A	-0.1464	0.0160
VAMP1	-0.1409	-0.4899
PLK1	-0.1373	-0.2386
CLIP2	-0.1344	0.0443
C11ORF95	-0.1308	-0.0153
HOXB2	-0.1291	-0.0217
SKIL	-0.1273	-0.2588
C1ORF159	-0.1262	0.0587
CPNE2	-0.1260	-0.2780

Gene symbol	RNA-seq Log2fc	Microarray Log2fc
TRIM56	-0.1254	-0.0852
ZNF789	-0.1099	-0.2293
MAP3K6	-0.1093	-0.3879
DUSP18	-0.1032	-0.0785
GPRC5A	-0.0917	-0.3093
POGZ	-0.0871	-0.2008
GRID1	-0.0618	-0.0160
KIAA1841	-0.0575	-0.2304
C11ORF80	-0.0572	-0.0432
CDKL5	-0.0340	-0.0927
BCAR1	-0.0327	0.0769
NBEAL1	0.0031	-0.0342
SMURF1	0.0070	-0.2964
ATXN7L3	0.0199	-0.1023
PJA2	0.0207	0.1329
JOSD1	0.0266	-0.0840
ZNF236	0.0289	-0.2232
RUSC1-AS1	0.0319	NaN
B9D2	0.0343	0.1092
SERPINB1	0.0389	0.1006
NTF4	0.0476	-0.0576
PC	0.0522	-0.0882
TTC7B	0.0540	-0.1838
TEP1	0.0569	-0.1652
IL18	0.0598	-0.0491
ZNF467	0.0827	-0.1821
RHBDL2	0.1029	-0.3546
NR6A1	0.1253	0.0828
AKR1C3	0.1303	0.1991
LMTK3	0.1473	-0.1184
ZNF414	0.1482	0.1164
ARNTL	0.1491	0.1382
FGD6	0.1807	-0.1192
ATL1	0.1844	0.0093
RAB4B	0.1853	0.1088
GAMT	0.1891	0.0339
C1ORF54	0.1904	0.0884
PPM1N	0.1944	-0.0093
OSBP2	0.2064	0.0039
BCL9	0.2300	0.2015
PDK2	0.2302	0.2762
GBP3	0.2435	0.1856
MN1	0.2456	-0.0309
GRHL1	0.2493	0.0130
C16ORF74	0.2744	-0.0772
PLK5	0.2833	NaN

Gene symbol	RNA-seq Log2fc	Microarray Log2fc
TRPM4	0.2854	0.0286
IL1R1	0.2901	-0.3124
SPOCD1	0.3008	0.0422
GAS2L2	0.3041	0.0057
SPTB	0.3177	0.0250
S100A8	0.3244	0.0062
WNT9A	0.3319	0.2649
NYNRIN	0.3416	-0.0779
IRAK2	0.3473	-0.0873
TP53INP1	0.4145	0.3892
BCO2	0.4552	0.1554
PCDHB4	0.5020	-0.1927
PDZD2	0.5050	0.1840
RNF152	0.5512	0.3983
GJB5	0.5583	0.0259
PODN	0.5583	0.0569
MMP1	0.5802	-0.0824
FGD2	0.5823	-0.0241
FBXO27	0.5976	0.0013
PTK2B	0.7125	0.0954
ADHFE1	0.8411	0.1434
CLSTN3	0.8906	0.1589
RNASE4	0.9100	NaN
GALNTL6	0.9851	0.1192
DSG1	1.0930	0.0092
LINC00173	1.3886	NaN
OR2A7	1.4328	0.0467
PRSS33	1.5150	-0.0032
S100A9	1.6366	0.3825

**P MYC Overexpression Up**  
(Figure 4.8D)

Gene symbol	RNA-seq Log2fc	Microarray Log2fc
C5ORF46	-1.5389	-0.0506
HS3ST3A1	-0.8786	0.0682
MMP16	-0.7549	-0.5514
FOS	-0.7177	-0.7646
TTYH2	-0.6935	0.1712
ABCC6P2	-0.6418	0.0282
LYPD6	-0.4973	-0.6607
VWA3B	-0.4850	0.2015
ENDOU	-0.4850	-0.0948
SLC6A15	-0.4850	0.0247
PLA2G4A	-0.4850	-0.0959

Gene symbol	RNA-seq Log2fc	Microarray Log2fc
DLL3	-0.4823	0.0436
ICAM5	-0.4753	-0.0290
DUSP2	-0.4185	-0.2427
TSFM	-0.3989	-0.0300
DPYSL5	-0.3694	-0.2777
MMP25	-0.3300	-0.0974
CHKA	-0.2727	-0.4210
HSPBAP1	-0.2609	-0.0252
EXOSC9	-0.2572	-0.2290
HS6ST2	-0.1768	0.0528
UNG	-0.1760	-0.1975
CALML4	-0.1454	-0.2407
PCOLCE2	-0.1260	-0.2078
NR2C2AP	-0.0963	0.0474
SCFD2	-0.0942	-0.2450
DGKD	-0.0793	-0.0769
CAPS	-0.0779	-0.0180
GLRX5	-0.0750	0.0777
TAF5	-0.0726	-0.1937
EEF1E1	-0.0707	0.0958
SLC29A4	-0.0664	0.1032
HLF	-0.0581	-0.0725
TGFBRAP1	-0.0379	-0.0429
TFB2M	-0.0370	0.0942
NIP7	-0.0307	-0.0261
ZNF549	-0.0286	0.2169
CD3EAP	-0.0286	-0.1372
JPH1	-0.0267	-0.3923
LAS1L	-0.0236	-0.0510
RRP15	-0.0230	-0.0453
TRNAU1AP	-0.0187	-0.1729
PRMT3	-0.0183	0.1218
SLC25A22	-0.0171	0.1185
PIM2	-0.0159	0.1828
UTP15	-0.0154	0.1132
LRRC61	-0.0118	-0.0216
DCAF4	-0.0101	-0.2984
TMEM97	-0.0076	-0.0572
GPD1L	-0.0011	-0.2330
MTFP1	0.0010	0.2368
ISOC2	0.0036	0.2281
NOC4L	0.0135	-0.0270
TMEM201	0.0179	0.0468
SAC3D1	0.0200	0.0770
PPRC1	0.0224	-0.1404
MRT04	0.0229	-0.1146

Gene symbol	RNA-seq Log2fc	Microarray Log2fc
CDCA7L	0.0242	-0.0968
FARSB	0.0273	-0.0044
RPP40	0.0298	0.1495
NDUFAF4	0.0308	0.4302
DPP7	0.0344	-0.0727
IPO4	0.0412	-0.0981
TSSC4	0.0513	0.2783
METTL1	0.0603	0.1571
DDX10	0.0607	-0.0582
AGPAT3	0.0616	0.0504
MPHOSPH10	0.0620	0.2799
ABCC4	0.0632	0.0320
PUS3	0.0660	0.1685
FOXRED2	0.0666	-0.2184
SLC25A15	0.0778	-0.1827
ESF1	0.0814	-0.0513
CCDC124	0.0819	-0.0565
WDR83	0.0872	0.0988
HSPE1	0.0880	0.3452
HDLBP	0.0882	-0.0633
ZNF593	0.0922	0.2665
PUS7	0.0949	-0.0336
SMYD5	0.0966	0.2535
DCTPP1	0.0980	0.3075
C20ORF27	0.1011	-0.0527
QTRT1	0.1024	-0.2774
SHISA9	0.1039	-0.2639
UTP14A	0.1046	0.1344
C19ORF48	0.1119	0.2212
PUS1	0.1126	0.0668
NOP14	0.1220	-0.0796
MON1A	0.1252	0.2631
ERVMER34-1	0.1254	NaN
NOL6	0.1259	0.0438
OAF	0.1263	-0.0134
DIS3L	0.1313	-0.1467
GEMIN5	0.1369	0.0739
PIGW	0.1388	0.1040
REPIN1	0.1393	0.1179
QSOX2	0.1393	-0.0463
RRP9	0.1406	0.0708
WDR4	0.1407	-0.0054
WDR43	0.1409	0.1714
DENND2D	0.1432	0.0745
USE1	0.1480	0.1814
RRP12	0.1525	0.2270

Gene symbol	RNA-seq Log2fc	Microarray Log2fc
PUS7L	0.1551	0.2430
PRR3	0.1571	-0.0778
PMM2	0.1586	0.3164
SSSCA1	0.1608	0.1889
TRMT1	0.1610	0.0546
PES1	0.1618	0.2179
MBLAC2	0.1646	-0.1472
AMPD2	0.1687	-0.0374
AHSA1	0.1713	0.2634
POLR1E	0.1884	0.1518
NLE1	0.1921	0.1311
EXOSC5	0.1974	0.1879
RPTOR	0.2015	0.0799
C12ORF66	0.2023	0.1541
TMEM132B	0.2089	0.0121
GLS2	0.2177	0.0589
PRR7	0.2195	0.0477
RPUSD4	0.2254	-0.0754
TOP1MT	0.2267	0.0967
LHX6	0.2314	-0.0959
PLD6	0.2366	0.1763
SLC19A1	0.2467	0.1146
MNT	0.2496	0.2485
CHCHD10	0.2533	0.3555
ZNF420	0.2554	0.1210
CD320	0.2598	0.0602
ALG3	0.2600	0.3876
UBIAD1	0.2690	0.2237
SORD	0.2755	0.1809
SLC25A26	0.2757	0.2588
KLRG2	0.2800	0.1898
B3GNTL1	0.2898	0.0438
CGREF1	0.3286	0.1350
CAMKMT	0.3287	NaN
SLC27A2	0.3365	0.3957
KLF16	0.3443	0.2695
TAF4B	0.3569	0.1366
P2RX5	0.3652	0.0542
ECE2	0.4124	0.2663
CCDC78	0.4301	-0.1835
IL17D	0.4805	0.0247
DGAT2	0.5336	0.2182
ZNF215	0.5583	0.0760
OR7C1	0.5583	-0.0559
ANO2	0.5583	0.1702
ITPR1	0.6051	0.3065

Gene symbol	RNA-seq Log2fc	Microarray Log2fc
GP2	0.9774	0.0312
HSPA6	1.0187	-0.0699

**Additional Tables 4.3 (A-P).** List of gene symbols and log2 fold change (fc) values for the heatmaps in Figures 4.6, 4.7, and 4.8. Genes are shown in the order displayed in the indicated figures, sorted by ascending RNA-seq fold change values. NaN indicates that microarray data are unavailable for this gene.

## Chapter 5: The Interaction Between ELF5 and DNA-PKcs

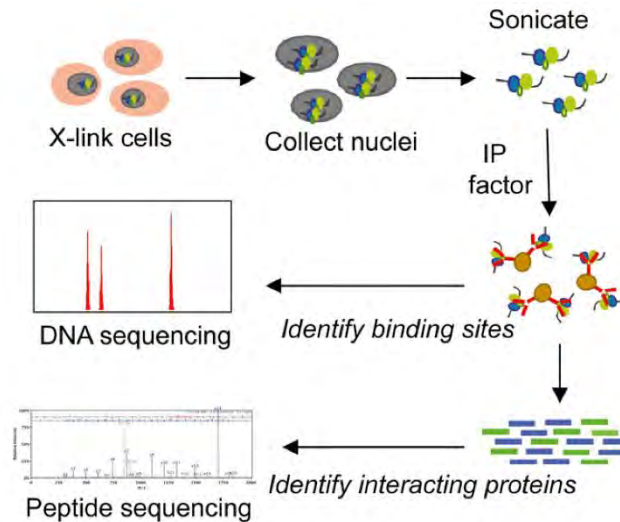
### Introduction

Interactions with other transcription factors, co-factors and post-translational modifiers are essential to the function of many transcriptional regulators. These interactions can vary according to cell type, external signals and genomic binding sites, mediating highly specific and finely-tuned alterations in gene expression. Although almost universally required for transcription factor action, no ELF5-interacting proteins have yet been identified in human breast or breast cancer cells. The aim of this study was to determine the human breast cancer ELF5 interactome, in order to identify potential mechanisms of ELF5 action and regulation.

The human interactome as a whole is estimated to contain 130,000 binary interactions, most of which are yet to be defined (Venkatesan *et al.*, 2008). There are a number of methods that may be used to unravel this complex protein interaction network. In general, these may be divided into methods that examine an interaction between two known proteins and methods that examine all interactions between a single known protein and its unknown partners. Examples in the first category include co-immunoprecipitation and proximity ligation assays (PLA) (both of which rely on antibodies specific to the proteins of interest) and additional immunofluorescence-based methods such as fluorescence resonance energy transfer (FRET) and bimolecular fluorescence complementation (BiFC) (relying on the overexpression of fluorophore-tagged proteins). Methods to identify the unknown interactors of a specific protein include computational methods (for example, the overlap of binding motifs in ChIP-seq data sets), yeast two-hybrid studies and mass spectrometry (MS) approaches. Two examples of MS-based approaches are tandem affinity purification (TAP) and the recently described rapid immunoprecipitation of endogenous proteins (RIME) (reviewed in Miller *et al.*, 2015).

RIME was developed specifically for the study of protein complexes involving chromatin and transcription factors (Mohammed *et al.*, 2013; Mohammed *et al.*, 2016) and was selected for the initial study of ELF5 interactions in breast cancer. RIME incorporates a formaldehyde cross-linking step, followed by bead-bound antibody immunoprecipitation (IP) to capture the protein of interest along with cross-linked members of the complex. The beads then undergo several stringent wash steps and on-bead tryptic digestion, releasing peptides from the protein complexes that are then

analysed by mass spectrometry (Figure 5.1). RIME has previously been used successfully in MCF7 breast cancer cells to identify proteins interacting with oestrogen receptor (Mohammed *et al.*, 2013), FOXA1 (Jozwik *et al.*, 2016) and lemur tyrosine kinase 3 (LMTK3) (Xu *et al.*, 2015), establishing the utility of the RIME method in this cell line.



**Figure 5.1: Rapid Immunoprecipitation of Endogenous Protein (RIME) purifies cross-linked transcriptional complexes**

**Figure reproduced with permission from Cell Press (Mohammed *et al.*, 2013)**

The cells are cross-linked using formaldehyde to stabilise protein complexes, followed by nuclear lysis and sonication (fragmenting DNA to 200-600 bp lengths). An immunoprecipitation is performed using an antibody against the protein of interest. After this step, the sample can be used for either RIME (to identify interacting proteins) or ChIP-seq (to identify DNA binding sites of the protein of interest). For RIME, on-bead tryptic digestion of bound proteins is followed by mass spectrometry analysis of peptides.

Due to the high level of redundancy in the ETS family DNA-binding domain, co-operative interactions are particularly important for specific ETS factor activity. Examples of proteins interacting with ETS factors include other transcription factors (such as ETS1 with RUNX1 and PAX5, discussed in Chapter 1, and ETV2 with FOXC2, discussed in Chapter 4), transcriptional co-factors (for example, the histone acetyltransferase CREBBP) and post-translational modifying enzymes (for example, the phosphorylation of ETS1 by CaMKII and MAPK1, discussed in Chapter 1). Many interactions involve the ETS DNA-binding domain and immediately adjacent regions (reviewed in Sharrocks, 2001). The interaction between ETS1 and PAX5, for example, occurs via the ETS domain, resulting in structural alterations that modify protein-DNA

contacts and facilitate binding to a non-canonical site (Garvie *et al.*, 2001). In addition, a number of ETS factor interactions involve the Pointed domain, which is present in a subset of ETS factors including ELF5. The ETS1 Pointed domain, for example, contains a docking site for MAPK1, which phosphorylates key residues N-terminal to the Pointed domain to enhance ETS1 transcriptional activity (reviewed in Garrett-Sinha, 2013; Hollenhorst *et al.*, 2011). In other family members, the Pointed domain interacts with repressive co-factors. The ETV6 (or TEL) Pointed domain, for example, interacts with the co-repressor SIN3A, driving the oncogenic activity of the ETV6-RUNX1 fusion protein in leukaemia (Fenrick *et al.*, 1999). The region between the Pointed and ETS domains is also the site of several important protein interactions, including the interaction between ETV6 and the co-repressor NCOR1 (Guidez *et al.*, 2000) and the interaction between ETS1 and the co-activator CREBBP (Yang *et al.*, 1998). The protein interaction interfaces of ETS factors have been identified as potential therapeutic targets in cancer (Cooper *et al.*, 2014), further emphasising the importance of understanding the contribution of protein interactions to ETS factor function.

Despite the importance of co-operative interactions to specific ETS factor function, there are very few published studies that have characterised ELF5 protein interactions. In one recent study, Elf5 (tagged with 3xFlag tag) was overexpressed in mouse trophoblastic stem cell (TSC) lines, immunoprecipitated and analysed by mass spectrometry, resulting in the identification of 109 potential ELF5-interacting proteins (Latos *et al.*, 2015). These included various transcription factors, such as eomesodermin (Eomes), transcription factor AP-2 gamma (Tfap2c), grainyhead-like 2 (Grhl2) and runt-related transcription factor 1 (Runx1), and chromatin modifiers, such as the SWI/SNF chromatin remodeller Smarca5, bromodomain PHD finger transcription factor (Bptf) and lysine demethylase 1A (Kdm1a). Interestingly, the level of Elf5 protein expression was shown to be an essential determinant in the preferential interaction with Eomes (promoting the stem cell state) or Tfap2c (promoting trophoblast differentiation), demonstrating the importance of differential Elf5 interactions in the regulation of cell fate. Another recent study suggested that ELF5 may interact with androgen receptor (AR) in human prostate cancer cells (Li *et al.*, 2017), although this has yet to be validated with additional methods. To date, however, there have been no global studies in any human cells to identify ELF5-interacting proteins and, similarly, there are no known post-translational modifications of ELF5 that may regulate its transcriptional function. This chapter presents the first use of RIME to identify novel ELF5-interacting proteins in human breast cancer cells.

Among the proteins identified was DNA-dependent protein kinase catalytic sub-unit (DNA-PKcs). As introduced in Chapter 1, this protein has diverse roles in DNA repair, mitosis, telomere maintenance, metabolism, immunity, hormone signalling, and transcriptional regulation. In cancer, the DNA repair functions of DNA-PKcs increase resistance to DNA-damaging cancer treatments such as radiotherapy and chemotherapy, providing the rationale for the clinical development of DNA-PKcs inhibitors (Velic *et al.*, 2015). DNA-PKcs has also been shown to regulate the transcriptional activity and expression of hormone receptors, including ER and AR, and to be essential for the transcriptional effects of highly expressed oncogenic ETS factors (Goodwin and Knudsen, 2014). The results presented in this chapter indicate an additional role for DNA-PKcs in regulating the function of ELF5 in breast cancer and represent an important contribution to the understanding of the transcriptional consequences of DNA-PKcs activity, as well as pharmacological inhibition, in breast cancer.



## Results

### Purification of ELF5-V5-associated proteins using RIME

MCF7-pHUSH-ELF5-Isoform2-V5 (MCF7-ELF5-V5) cells were treated with doxycycline to induce ELF5-Isoform2-V5 (ELF5-V5) expression, which was purified from  $5-10 \times 10^7$  cells using RIME (Mohammed *et al.*, 2013; Mohammed *et al.*, 2016). Briefly, the cells were cross-linked using formaldehyde and ELF5-V5 was immunoprecipitated using a combination of ELF5 and V5 antibodies bound to magnetic beads. The extracted proteins then underwent on-bead tryptic digestion and analysis by mass spectrometry (Figure 5.1). A total of five ELF5-V5 RIME replicates with parallel IgG controls were performed. The first replicate (RIME 1) was performed by collaborators at Cancer Research UK (Cambridge, UK) and four subsequent replicates (RIME 2-5) were performed locally in conjunction with the Australian Proteome Analysis Facility (Macquarie University, Sydney, Australia). The numbers of proteins identified in each replicate are summarised in Table 5.1.

**Table 5.1: Numbers of proteins identified in ELF5-V5 and IgG control RIME replicates**

	IgG	ELF5	ELF5-specific
RIME 1	316	527	246
RIME 2	39	276	113
RIME 3	21	144	50
RIME 4	27	61	14
RIME 5	30	117	24
Total	433	1125	447
Total unique	346*	637	341**
Total unique 2/5 reps			74**
Total unique 3/5 reps			21
Total unique 4/5 reps			8
Total unique 5/5 reps			3

Numbers of proteins identified in the IgG controls (col 2) and ELF5-V5 experiments (col 3) for each RIME replicate. Total number of proteins = column sum, while total *unique* proteins = column sum with duplicates (those proteins identified in more than experiment) removed. The total unique value for the IgG experiments (346) represents the list of IgG-identified proteins used to create the ELF5-specific protein lists by subtraction (column 4). The ELF5-specific proteins therefore represent those proteins identified in the ELF5-V5 RIME that were not identified in any IgG RIME experiment. Removal of duplicate proteins results in a total of 341 ELF5-specific proteins. Of these, 74 were identified in at least 2 of the 5 ELF5-V5 RIME replicates, 21 were identified in 3 replicates, 8 were identified in 4 replicates, and 3

were identified in all 5 replicates. The 74 proteins (or 73 excluding ELF5) identified in at least 2 replicates are detailed in Table 5.2. \*Combined IgG list used as the control for analysis of ELF5 replicates. \*\* Protein sets used for downstream analyses.

As can be seen from this table, there was a large variation in the numbers of proteins identified in individual replicates. By some margin, the largest numbers of proteins were found in replicate 1, with 527 proteins identified in the ELF5-V5 RIME and 316 proteins identified in the IgG control. In general, RIME experiments are expected to identify 300-900 proteins, of which 5-10% will be specific interactors. The lower numbers of proteins identified in replicates 2-5 may indicate technical issues such as inefficient antibody coupling or washing, excessive cross-linking or, in the case of experiment 5, insufficient cell number (Mohammed *et al.*, 2016). In all replicates, however, ELF5 was identified as one of the top-ranking proteins by Mascot score, ranging from rank 1-14 following removal of non-specific interactors. This indicates successful, although variable, ELF5-V5 purification of interacting proteins in all replicates.

A stringent list of non-specific interactors was generated using all proteins identified in any IgG replicate (346 proteins). As can be seen in Figure 5.2A, many of these non-specific proteins were unique to replicate 1. Only ELF5-V5 proteins that did not occur in this combined list were considered for further analysis. These numbers are summarised in the “ELF5-specific” column of Table 5.1 and ranged from only 14 proteins in replicate 4 to 246 proteins in replicate 1.

The overlap between ELF5-V5 replicate experiments following removal of non-specific interactors is shown in Figure 5.2B. Three proteins were identified in all five ELF5-V5 replicates and none of the IgG controls. These were ELF5 itself, DNA-dependent protein kinase catalytic subunit (DNA-PKcs), and protein transport protein SC16A. While proteins identified in fewer replicates are more likely to represent false-positives, the limited numbers of proteins identified in several replicates indicates they may have limited sensitivity. Therefore, proteins identified in at least 2 of the 5 replicates were also considered to be potential ELF5 interactors. This approach resulted in a total of 73 candidate ELF5-interacting proteins (excluding ELF5), summarised in Table 5.2.

Several proteins known to interact with DNA-PKcs were identified, including XRCC5/Ku80 (3/5 replicates), DNA topoisomerase 2-beta (TOP2B, 2/5 replicates), DNA topoisomerase 2-alpha (TOP2A, 1/5 replicates) and poly(ADP-ribose) polymerase 1 (PARP1, 1/5 replicates). The second Ku sub-unit (XRCC6/Ku70) was identified in 2/5 replicates but was also found in one IgG control experiment and was therefore excluded.

**Table 5.2: ELF5-interacting proteins identified by RIME**

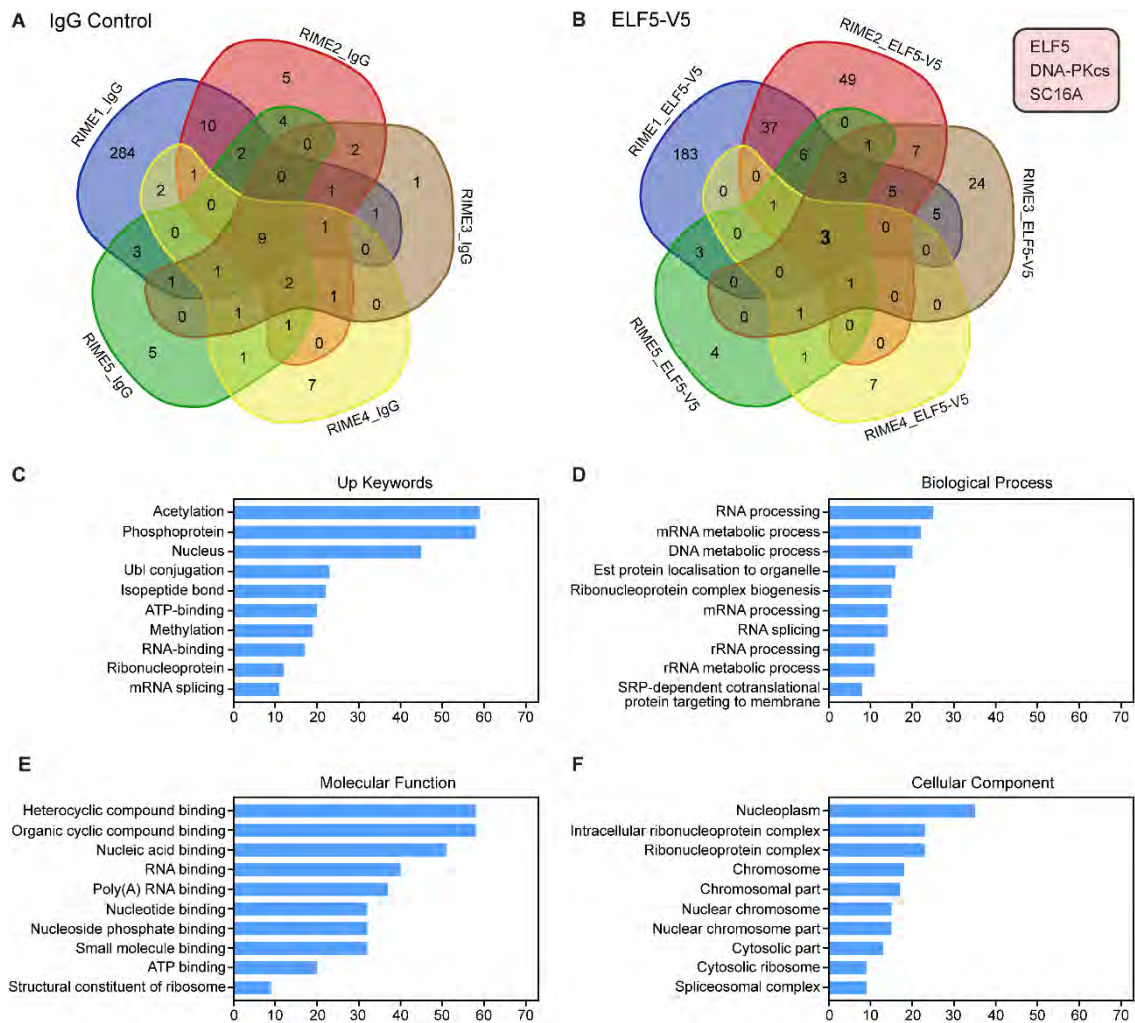
Uniprot ID	Uniprot Protein Name	Unique Peptides					Mascot Score				
		1 (%)	2	3	4	5	1	2	3	4	5
5 of 5 replicates (3)											
ELF5_HUMAN	ETS-related transcription factor Elf-5	4 (15.5)	4	2	4	5	490	587	146	213	365
PRKDC_HUMAN	DNA-dependent protein kinase catalytic subunit	15 (4.8)	6	1	1	1	239	128	42	47	51
SC16A_HUMAN	Protein transport protein Sec16A	14 (12.5)	20	7	1	2	159	1244	278	85	202
4 of 5 replicates (5)											
RL13A_HUMAN	60S ribosomal protein L13a	4 (22.7)	1	2		1	80	51	118		73
RL18A_HUMAN	60S ribosomal protein L18a		2	2	1	1		85	123	57	54
RL30_HUMAN	60S ribosomal protein L30	2 (17.4)	2	2		1	29	121	78		84
TRI25_HUMAN	E3 ubiquitin/ISG15 ligase TRIM25	2 (3.5)	1		1	1	148	43		51	55
U2AF1_HUMAN	Splicing factor U2AF 35 kDa subunit	2 (10.8)	2	1		1	56	111	68		62
3 of 5 replicates (13)											
ATPA_HUMAN	ATP synthase subunit alpha, mitochondrial	1 (4.0)	1			1	35	69			90
H12_HUMAN	Histone H1.2	7 (17.8)	2	1			479	126	98		
H2B1C_HUMAN	Histone H2B type 1-C/E/F/G/I			2	3	2			75	96	89
PAIRB_HUMAN	Plasminogen activator inhibitor 1 RNA-binding protein	2 (7.8)	1			1	62	81			84
PPM1G_HUMAN	Protein phosphatase 1G	8 (25.5)	3	1			186	100	48		
PUR6_HUMAN	Multifunctional protein ADE2	2 (10.4)	1			1	87	62			109
RA1L2_HUMAN	Heterogeneous nuclear ribonucleoprotein A1-like 2		2	1		2		160	48		134
RCC1_HUMAN	Regulator of chromosome condensation	4 (15.7)	1			1	27	129			45
RS27A_HUMAN	Ubiquitin-40S ribosomal protein S27a	3 (27.6)	4	1			84	276	154		
TCPE_HUMAN	T-complex protein 1 subunit epsilon	1 (4.3)	1			1	61	87			53
TRI33_HUMAN	E3 ubiquitin-protein ligase TRIM33	3 (2.9)	3	1			138	102	64		
U2AF2_HUMAN	Splicing factor U2AF 65 kDa subunit	1 (2.1)	1			1	28	43			98
XRCC5_HUMAN	X-ray repair cross-complementing protein 5	13 (30.1)	2	1			166	50	45		
2 of 5 replicates (53)											
ACL6A_HUMAN	Actin-like protein 6A	1 (4.4)	1				21	38			
ACTG_HUMAN	Actin, cytoplasmic 2		6	9				744	465		
ATPB_HUMAN	ATP synthase subunit beta, mitochondrial		1	2				77	82		

Uniprot ID	Uniprot Protein Name	Unique Peptides					Mascot Score				
		1 (%)	2	3	4	5	1	2	3	4	5
C2TA_HUMAN	MHC class II transactivator		1	1				38	42		
CBX1_HUMAN	Chromobox protein homolog 1	1 (17.8)	1				181	130			
CCAR2_HUMAN	Cell cycle and apoptosis regulator protein 2	11 (21.0)	1				358	55			
CHD4_HUMAN	Chromodomain-helicase-DNA-binding protein 4	1 (0.9)	1				27	62			
CLIC1_HUMAN	Chloride intracellular channel protein 1	1 (7.5)	2				48	116			
CO3A1_HUMAN	Collagen alpha-1(III) chain	2 (5.2)	3				91	117			
COPB_HUMAN	Coatomer subunit beta	2 (5.1)	1				24	83			
DDX46_HUMAN	Probable ATP-dependent RNA helicase DDX46	1 (1.9)	1				20	45			
DUS23_HUMAN	Dual specificity protein phosphatase 23	5 (37.3)	1				195	46			
DX39A_HUMAN	ATP-dependent RNA helicase DDX39A	2 (17.8)	1				171	55			
H10_HUMAN	Histone H1.0	1 (4.6)	1				23	58			
HNRC1_HUMAN	Heterogeneous nuclear ribonucleoprotein C-like 1				1	1				91	87
IF2A_HUMAN	Eukaryotic translation initiation factor 2 subunit 1	3 (11.8)		1			87		66		
IMA6_HUMAN	Importin subunit alpha-6	1 (2.6)	1				41	37			
LAP2A_HUMAN	Lamina-associated polypeptide 2, isoform alpha	4 (15.6)	1				178	59			
LASP1_HUMAN	LIM and SH3 domain protein 1	3 (21.8)	2				33	72			
LC7L2_HUMAN	Putative RNA-binding protein Luc7-like 2	2 (5.9)				1	38				63
MCM3_HUMAN	DNA replication licensing factor MCM3	2 (4.3)	2				60	115			
MCM5_HUMAN	DNA replication licensing factor MCM5	2 (3.4)	1				24	96			
MCM6_HUMAN	DNA replication licensing factor MCM6	1 (1.6)	1				28	48			
NOG1_HUMAN	Nucleolar GTP-binding protein 1	1 (2.7)	1				44	41			
NP1L1_HUMAN	Nucleosome assembly protein 1-like 1	2 (7.4)	1				65	89			
NPM3_HUMAN	Nucleoplasmin-3		1	1				133	71		
PRP19_HUMAN	Pre-mRNA-processing factor 19	3 (18.9)	1				114	44			
PSA4_HUMAN	Proteasome subunit alpha type-4	1 (3.8)	1				64	69			
PTBP2_HUMAN	Polypyrimidine tract-binding protein 2	16 (62.3)	4				1059	250			
RAVR1_HUMAN	Ribonucleoprotein PTB-binding 1	17 (43.9)	5				1478	260			
RAVR2_HUMAN	Ribonucleoprotein PTB-binding 2	12 (39.4)	6				505	271			
RL17_HUMAN	60S ribosomal protein L17	3 (15.8)		2			72		61		

Uniprot ID	Uniprot Protein Name	Unique Peptides					Mascot Score				
		1 (%)	2	3	4	5	1	2	3	4	5
RL35A_HUMAN	60S ribosomal protein L35a	2 (22.7)		1			49		43		
RL5_HUMAN	60S ribosomal protein L5	4 (16.2)	1				67	53			
RMXL1_HUMAN	RNA binding motif protein, X-linked-like-1		1	1				50	61		
RS12_HUMAN	40S ribosomal protein S12	2 (13.6)	1				34	98			
RS27L_HUMAN	40S ribosomal protein S27-like	1 (28.6)		1			32		55		
SC23A_HUMAN	Protein transport protein Sec23A		5	1				257	41		
SC23B_HUMAN	Protein transport protein Sec23B	2 (7.0)	8				26	419			
SRSF9_HUMAN	Serine/arginine-rich splicing factor 9	3 (19.0)	1				34	50			
SSRP1_HUMAN	FACT complex subunit SSRP1	3 (5.2)	2				53	122			
SYEP_HUMAN	Bifunctional glutamate/proline--tRNA ligase	2 (2.4)				1	27				45
SYFA_HUMAN	Phenylalanine--tRNA ligase alpha subunit	1 (2.8)	1				45	83			
TCP4_HUMAN	Activated RNA polymerase II transcriptional coactivator p15	2 (22.1)	2				30	98			
TCPG_HUMAN	T-complex protein 1 subunit gamma	6 (19.6)	1				133	62			
TCPH_HUMAN	T-complex protein 1 subunit eta	2 (5.7)				1	57				56
TCPQ_HUMAN	T-complex protein 1 subunit theta	2 (6.6)	1				56	115			
THOC4_HUMAN	THO complex subunit 4	4 (31.5)	1				97	56			
TLE1_HUMAN	Transducin-like enhancer protein 1		1	1				62	42		
TOP2B_HUMAN	DNA topoisomerase 2-beta	2 (3.8)	1				202	57			
UBF1_HUMAN	Nucleolar transcription factor 1	2 (6.0)	1				24	53			
VDAC2_HUMAN	Voltage-dependent anion-selective channel protein 2	2 (7.5)		1			66		48		
XPO2_HUMAN	Exportin-2	1 (3.0)	2				54	153			

**Table 5.2: ELF5-interacting proteins identified by RIME.** List of ELF5-specific proteins identified in at least 2 of 5 RIME replicates. The first two columns show the Uniprot database protein ID and full name. Column 3 shows the number of unique peptides identified for the protein in each of the RIME replicates (sub-columns 1-5). RIME replicate 1 (sub-column 1) also shows the coverage of the identified peptides as a percentage of the full protein sequence in parentheses. Grey shading indicates that no peptides were identified for the protein in the corresponding experiment. Column 4 shows the Mascot score for the protein in each of the RIME replicates (sub-columns 1-5).

Gene ontology (GO) analysis of the 73 proteins identified is shown in Figures 5.2C-F. The top 10 GO terms for each category reveal some general patterns, including involvement in DNA/RNA binding and RNA processing and localisation to the nuclear compartment, broadly consistent with components of transcriptional complexes.



**Figure 5.2: ELF5-V5 RIME purifies multiple proteins in MCF7 luminal breast cancer cells**

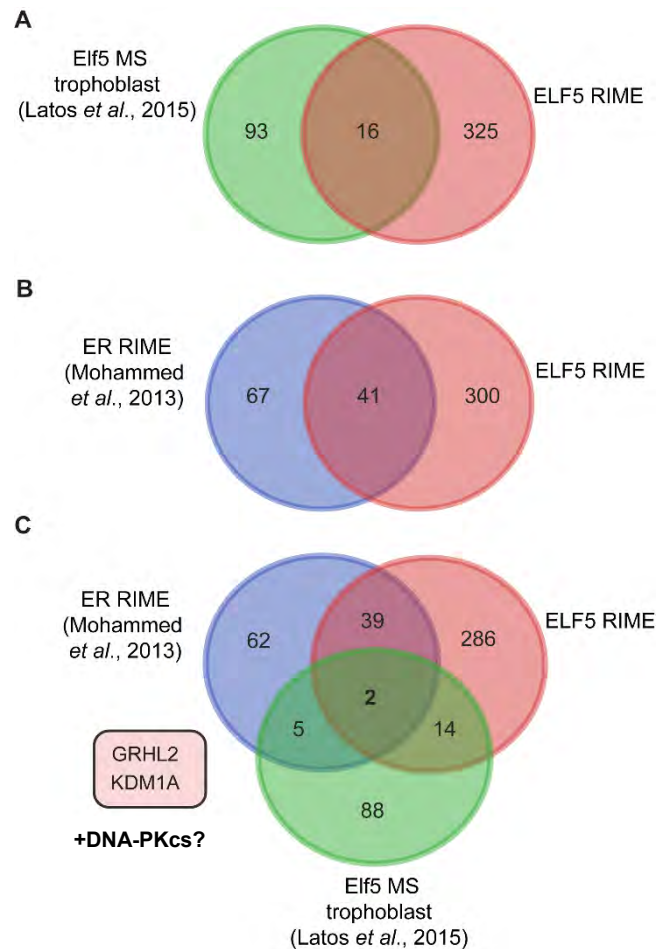
(A) Overlap of proteins identified in the IgG controls for RIME replicates 1-5, representing non-specific interactors. (B) Overlap of proteins identified in the ELF5-V5 RIME replicates 1-5 following removal of non-specific interactors. Three proteins (ELF5, DNA-PKcs and SC16A) were identified in all replicates. (C-F) Gene ontology analysis for the 73 proteins identified in at least 2 ELF5-V5 RIME replicates, including enriched keywords (C), biological process (D), molecular function (E) and cellular component (F).

### **Comparison of human ELF5, murine Elf5, ER, and FOXA1 interactomes**

Potential ELF5-interacting proteins (identified in any ELF5-V5 RIME replicate) were compared to proteins identified in the previously published Elf5 MS study in mouse trophoblast stem cells (Latos *et al.*, 2015). Of the 109 proteins identified in this study, 16 were also found in at least one ELF5-V5 replicate in MCF7 cells (Figure 5.3A and Table 5.3). This points to some potential ELF5 interactors that may be common between cell types and species. Some interesting common proteins identified included the transcription factor GRHL2, the lysine-specific demethylase KDM1A and the nucleosome remodelling protein bromodomain PHD finger transcription factor (BPTF). Each of these proteins were only identified in 1 of 5 ELF5-V5 RIME replicates, indicating the need for further validation of these candidates. DNA-PKcs and SEC16A, present in 5 of 5 ELF5-V5 RIME replicates, were not identified in this study.

Due to the known relationship between ER and ELF5 in breast cancer, ELF5-V5 RIME proteins were also compared with the stringent list of ER interactors in MCF7 cells identified by RIME (Mohammed *et al.*, 2013). Of the 108 ER-interacting proteins, 41 (38%) were also identified in at least 1 ELF5-V5 replicate (Figure 5.3B and Table 5.4). This included 8 confirmed ER interactors identified in previous studies (highlighted in bold Table 5.4). Interestingly, GRHL2 and KDM1A were again found in both ER and ELF5-V5 RIME, hinting at a potentially important role for these two proteins in both ER and ELF5 transcriptional regulation. These were the only 2 proteins identified in all 3 interactomes (Figure 5.3C). DNA-PKcs was excluded from the stringent list of ER interactors as it was identified in at least one IgG control experiment; however, this may be a false-negative, as DNA-PKcs has been shown in previous studies to interact with ER and to regulate its phosphorylation and transcriptional activity (Foulds *et al.*, 2013; Liu *et al.*, 2014; Medunjanin *et al.*, 2010b).

Finally, ELF5-V5 RIME proteins were compared to those proteins identified in FOXA1 RIME in MCF7 cells (Jozwik *et al.*, 2016). FOXA1 RIME identified 250 proteins, of which 48 (19%) were common to ELF5-V5 RIME (data not shown). GRHL2 was again common to both datasets, with GRHL2 ChIP-seq experiments revealing a high degree of overlap between GRHL2, FOXA1 (and novel FOXA1 interactor MLL3) binding. Therefore, GRHL2 and KDM1A were identified as potential ELF5 interaction candidates through these comparisons, in addition to the proteins DNA-PKcs and SEC16A that were identified in all ELF5-V5 RIME replicates.



**Figure 5.3: ELF5-V5 RIME proteins overlap with published mouse Elf5 and human ER interacting proteins**

(A) Overlap between ELF5-interacting proteins identified in at least one MCF7 ELF5-V5 RIME replicate and those identified in mouse trophoblast cells (Latos *et al.*, 2015).

(B) Overlap between ELF5-interacting proteins identified in at least one ELF5-V5 RIME replicate and ER-interacting proteins identified by ER RIME in MCF7 cells (Mohammed *et al.*, 2013).

(C) Overlap of the three interactomes, revealing GRHL2 and KDM1A as common interactors. DNA-PKcs, a likely false-negative in the ER RIME experiment, may also be in this group.



**Table 5.3: Common proteins identified in ELF5-V5 RIME (MCF7 cells) and Elf5 MS (mouse trophoblastic stem cells)**

Accession	Uniprot ID	Protein Name
Q12830	BPTF_HUMAN	Nucleosome-remodeling factor subunit BPTF
Q9UKW6	ELF5_HUMAN	ETS-related transcription factor Elf-5
Q6ISB3	GRHL2_HUMAN	Grainyhead-like protein 2 homolog
P16403	H12_HUMAN	Histone H1.2
P51610	HCFC1_HUMAN	Host cell factor 1
P20042	IF2B_HUMAN	Eukaryotic translation initiation factor 2 subunit 2
P52292	IMA1_HUMAN	Importin subunit alpha-1
O60341	KDM1A_HUMAN	Lysine-specific histone demethylase 1A
Q9UQ80	PA2G4_HUMAN	Proliferation-associated protein 2G4
O75400	PR40A_HUMAN	Pre-mRNA-processing factor 40 homolog A
Q02543	RL18A_HUMAN	60S ribosomal protein L18a
P62888	RL30_HUMAN	60S ribosomal protein L30
P46777	RL5_HUMAN	60S ribosomal protein L5
P62841	RS15_HUMAN	40S ribosomal protein S15
P50990	TCPQ_HUMAN	T-complex protein 1 subunit theta
Q02880	TOP2B_HUMAN	DNA topoisomerase 2-beta

**Table 5.4: Common proteins identified in ELF5-V5 and ER RIME (MCF7 cells)**

Accession	Uniprot ID	Protein Name
P12814	ACTN1_HUMAN	Alpha-actinin-1
Q9BTT0	AN32E_HUMAN	Acidic leucine-rich nuclear phosphoprotein 32 family member E
P45973	CBX5_HUMAN	Chromobox protein homolog 5
Q8N163	CCAR2_HUMAN	Cell cycle and apoptosis regulator protein 2
Q14839	CHD4_HUMAN	Chromodomain-helicase-DNA-binding protein 4
Q9UNE7	CHIP_HUMAN	E3 ubiquitin-protein ligase CHIP
P56545	CTBP2_HUMAN	C-terminal-binding protein 2
Q6NXG1	ESRP1_HUMAN	Epithelial splicing regulatory protein 1
P23771	GATA3_HUMAN	Trans-acting T-cell-specific transcription factor GATA-3
Q6ISB3	GRHL2_HUMAN	Grainyhead-like protein 2 homolog
P78347	GTF2I_HUMAN	General transcription factor II-I
O14929	HAT1_HUMAN	Histone acetyltransferase type B catalytic subunit
Q92769	HDAC2_HUMAN	Histone deacetylase 2
Q9BUJ2	HNRL1_HUMAN	Heterogeneous nuclear ribonucleoprotein U-like protein 1
O60341	KDM1A_HUMAN	Lysine-specific histone demethylase 1A
P42166	LAP2A_HUMAN	Lamina-associated polypeptide 2, isoform alpha
P25205	MCM3_HUMAN	DNA replication licensing factor MCM3
P33991	MCM4_HUMAN	DNA replication licensing factor MCM4
P33992	MCM5_HUMAN	DNA replication licensing factor MCM5
P55209	NP1L1_HUMAN	Nucleosome assembly protein 1-like 1
Q9ULU4	PKCB1_HUMAN	Protein kinase C-binding protein 1
P30041	PRDX6_HUMAN	Peroxiredoxin-6
O95758	PTBP3_HUMAN	Polypyrimidine tract-binding protein 3
Q16576	RBBP7_HUMAN	Histone-binding protein RBBP7
Q9BWF3	RBM4_HUMAN	RNA-binding protein 4
Q92785	REQU_HUMAN	Zinc finger protein ubi-d4
P42677	RS27_HUMAN	40S ribosomal protein S27
Q15637	SF01_HUMAN	Splicing factor 1
O75533	SF3B1_HUMAN	Splicing factor 3B subunit 1
Q15427	SF3B4_HUMAN	Splicing factor 3B subunit 4
Q969G3	SMCE1_HUMAN	SWI/SNF-related matrix-associated actin-dependent regulator of chromatin E1
Q9BZK7	TBL1R_HUMAN	F-box-like/WD repeat-containing protein TBL1XR1
P10599	THIO_HUMAN	Thioredoxin
Q86V81	THOC4_HUMAN	THO complex subunit 4
Q04724	TLE1_HUMAN	Transducin-like enhancer protein 1
Q9UPN9	TRI33_HUMAN	E3 ubiquitin-protein ligase TRIM33
Q9BRA2	TXD17_HUMAN	Thioredoxin domain-containing protein 17
Q93009	UBP7_HUMAN	Ubiquitin carboxyl-terminal hydrolase 7
P55060	XPO2_HUMAN	Exportin-2 (Exp2)
Q96MM3	ZFP42_HUMAN	Zinc finger protein 42 homolog
O75362	ZN217_HUMAN	Zinc finger protein 217

## DNA-PKcs expression and alterations in breast cancer

Due to the robust discovery of DNA-PKcs in all ELF5-V5 RIME replicates, as well as its known roles in transcriptional regulation, the decision was made to focus further studies on validation of the potential interaction between DNA-PKcs and ELF5. An initial screen was performed to examine the expression of *PRKDC* (the gene encoding DNA-PKcs) across various normal tissues and cancers in samples from The Cancer Genome Atlas (TCGA) (Figure 5.4A). To minimise the use of multiple names, the *PRKDC* gene will be referred to hereafter as *DNA-PKcs* (italicised).

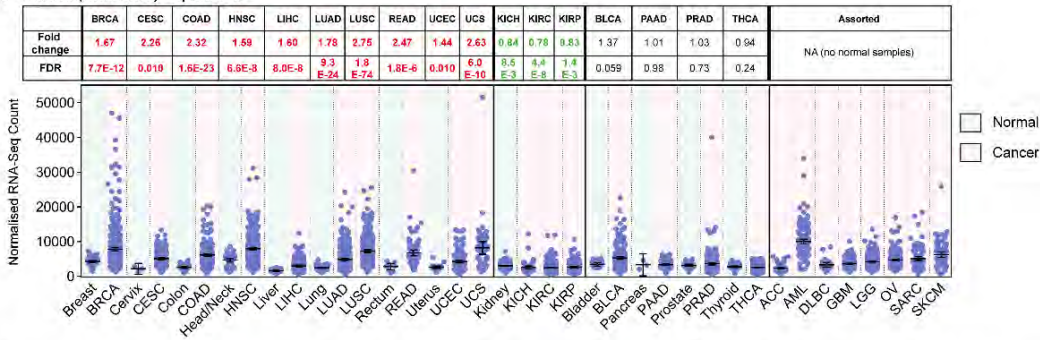
This analysis demonstrated a significant mRNA upregulation of *DNA-PKcs* in multiple cancer types, including breast cancer. Interestingly, there was a small but significant downregulation of *DNA-PKcs* in all three types of kidney carcinoma. Breast cancer ranked 4th (out of 25 cancer types) for mean *DNA-PKcs* expression, behind acute myeloid leukaemia, uterine carcinosarcoma and head/neck carcinoma. No normal tissues had any samples with *DNA-PKcs* expression over 10,000, while 21 of 25 cancer types had at least one sample with expression over 10,000. In the breast cancer set, 19.6% (101/515) of samples had expression over 10,000 and 2.9% (15/515) had expression over 20,000. Breast cancer had the largest proportion of samples with *DNA-PKcs* expression over 20,000 of any cancer type.

*DNA-PKcs* expression was then analysed in each of the molecular subtypes of breast cancer. As shown in Figure 5.4B, the expression of *DNA-PKcs* was significantly increased compared to the normal breast in all molecular subtypes of breast cancer (with the exception of normal-like) in the TCGA RNA-seq dataset. The fold change increase was greatest in the basal-like subtype (2.34-fold) and smallest in the luminal A subtype (1.42-fold). A subset of these samples also had quantitative proteomic data available (Mertins *et al.*, 2016), which demonstrated a similar pattern of upregulation (Figure 5.4C). There was a weak positive correlation between RNA and protein quantities measured in samples from the same tumour (Pearson  $r=0.26$ , FDR = 0.03), slightly below the median correlation ( $r=0.39$ ) for the entire dataset (Mertins *et al.*, 2016) (Figure 5.4D).

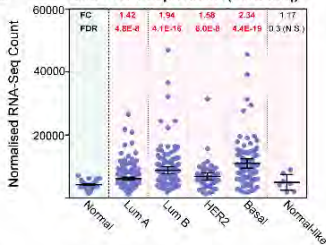
Next, the correlation between *DNA-PKcs* and *ELF5* expression in the different molecular subtypes was examined (Figure 5.4E). There was no correlation between expression of these two genes in any molecular subtype, with *ELF5* expression strongly influenced by molecular subtype (consistent with previous studies) and *DNA-PKcs* expression varying widely within each subtype. There was a weak positive correlation between *DNA-PKcs* and *ELF5* expression in the normal breast (Spearman

$r = 0.39$ ,  $p = 0.0016$ ). Similarly, there was no correlation between *DNA-PKcs* and *ESR1* expression in any molecular subtype of breast cancer or in the normal breast, despite previous studies demonstrating a positive-feedback relationship between these two proteins (Figure 5.4F).

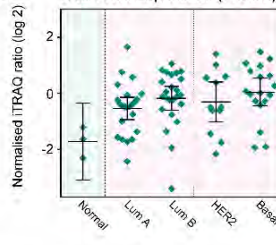
#### A *PRKDC* (DNA-PKcs) Expression



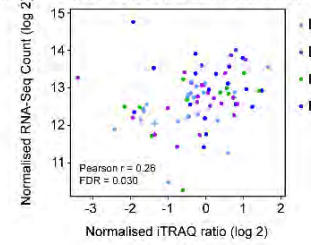
#### B *DNA-PKcs* Expression (RNA-Seq)



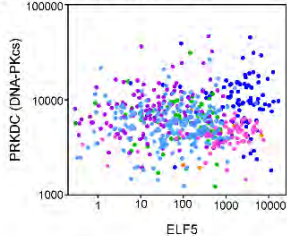
#### C *DNA-PKcs* Expression (Protein)



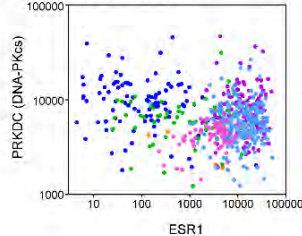
#### D *DNA-PKcs* RNA/Protein Correlation



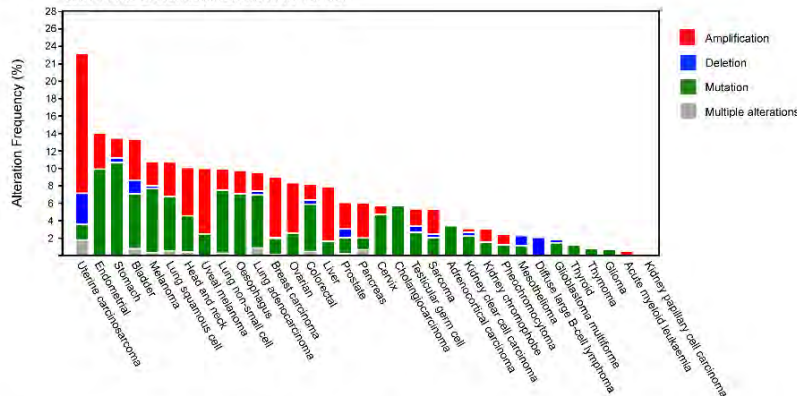
#### E *ELF5* and *DNA-PKcs*



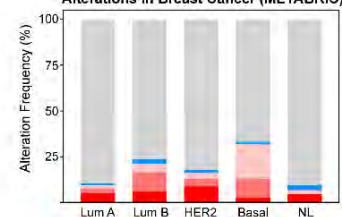
#### F *ESR1* and *DNA-PKcs*



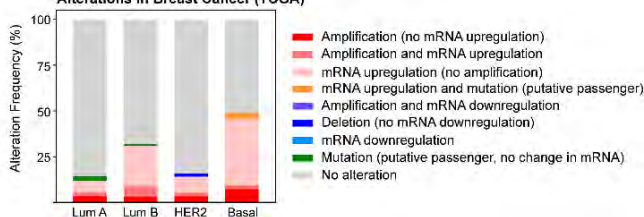
#### G *DNA-PKcs* Genomic Alterations in Cancer



#### H Subtype-Specific *DNA-PKcs* Alterations in Breast Cancer (METABRIC)



#### I Subtype-Specific *DNA-PKcs* Alterations in Breast Cancer (TCGA)



### Figure 5.4: DNA-PKcs is significantly altered in cancer

(previous page)

(A) *DNA-PKcs* (also known as *PRKDC*) gene expression from TCGA for 25 cancer types (pink background), with normal tissue comparisons (green background) where available. Plotted values represent individual TCGA RNA-sequencing samples (normalised count) and error bars the mean with 95% confidence interval. TCGA cancer acronyms are used (see Table 3.1). Fold change and False Discovery rate (FDR) from limma voom analysis are shown, with green values in bold indicating significant downregulation and red values in bold indicating significant upregulation compared to normal (FDR<0.05). (B) *DNA-PKcs* gene expression (normalised RNA-seq counts) for normal breast and breast cancer subtypes (n=585), with fold change (FC) and False Discovery Rate (FDR) from limma voom analysis as above. Error bars represent the mean with 95% confidence interval. (C) *DNA-PKcs* protein levels for a subset of 77 subtyped TCGA breast cancer samples (data from Mertins *et al.*, 2016). Isobaric peptide labelling (Isobaric Tags for Relative and Absolute Quantification or iTRAQ) was used to quantify protein levels, which are expressed as the normalised log<sub>2</sub> iTRAQ ratio. (D) Correlation between *DNA-PKcs* RNA and protein levels for the 77 TCGA samples with matched data. Molecular subtypes are indicated by colour. (E) Correlation between *DNA-PKcs* and *ELF5* RNA expression (normalised RNA-seq count) in the subtyped TCGA breast cancer cohort, with molecular subtypes indicated by colour (n=585). (F) Correlation between *DNA-PKcs* and *ER* RNA expression (normalised count) in the subtyped TCGA breast cancer cohort. (G) Frequency of *DNA-PKcs* genomic alterations (amplifications, deletions, mutations) in 33 different TCGA cancer types. (H) Combined *DNA-PKcs* gene expression changes and genomic alterations in breast cancer molecular subtypes for the METABRIC dataset; no mutation data is available for *DNA-PKcs*. (I) Combined *DNA-PKcs* gene expression changes and genomic alterations in breast cancer molecular subtypes for the subtyped TCGA breast cancer cohort.

Genomic alterations in *DNA-PKcs* were also investigated in multiple TCGA cancer datasets using cBioPortal (Cerami *et al.*, 2012; Gao *et al.*, 2013) (Figure 5.4G). Amplifications and mutations were the most frequently detected genomic alterations, while deletions were relatively rare. Uterine cancers (including carcinosarcoma and endometrial carcinoma) had the highest rates of combined *DNA-PKcs* alterations, occurring in 23.2% and 14.0% of samples respectively. Uterine carcinosarcoma also had the highest rate of *DNA-PKcs* amplification (16.1% of samples, n=9) and deletion (3.6% of samples, n=2). The highest rate of *DNA-PKcs* mutation occurred in stomach adenocarcinoma (10.7% of samples, n=42). Acute myeloid leukaemia had an extremely low rate of genomic alterations (only a single amplification out of 188 cases) despite having the highest average *DNA-PKcs* expression level of any cancer.

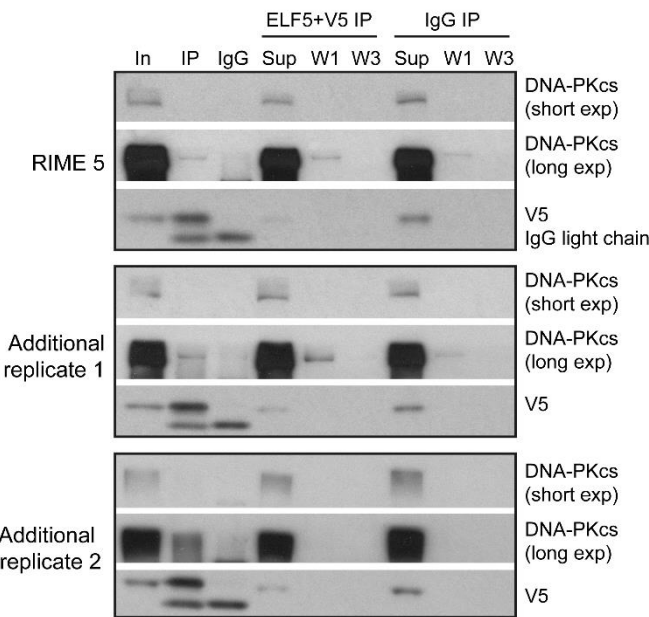
In breast cancer, DNA-PKcs was altered in 9.0% (87/963) of cases. The main alteration in breast cancer was DNA-PKcs amplification (7.0% of samples, n = 67), while mutation occurred in 1.9% of samples (n=18) and only a single case of deletion was identified. Breast cancer had the third-highest rate of DNA-PKcs amplification, exceeded only by uterine carcinosarcoma and uveal melanoma.

Finally, gene expression and genomic alteration data were combined for subtyped breast cancer samples in the METABRIC (n = 1979) (Curtis *et al.*, 2012; Pereira *et al.*, 2016) and TCGA (n = 508) datasets (Figures 5.4H and 5.4I). Up- or down-regulation of mRNA was defined as a z-score of more than  $\pm 2.0$  compared to the expression distribution for samples diploid for DNA-PKcs. In both datasets, the basal-like subtype had the greatest level of combined DNA-PKcs changes (33.1% and 49.0% of samples respectively), while the luminal A subtype had the lowest (10.4% and 14.4% respectively). The normal-like subtype in the METABRIC dataset also had a low level of changes (9.5%) but was excluded in the TCGA dataset due to the small number of samples. Upregulation of *DNA-PKcs* mRNA ( $\pm$  genomic amplification) was the most common change in DNA-PKcs in all subtypes, while downregulation was uncommon. In some cases, mRNA upregulation and amplification occurred together (medium red), however a significant proportion of cases demonstrated increased expression in the absence of amplification (light red) and, to a lesser extent, amplification in the absence of increased expression (dark red).

In summary, these large-scale and subtype-specific studies demonstrate that DNA-PKcs is commonly altered in breast cancer, at both the genomic level (primarily amplification and mutation) and expression level (upregulation). DNA-PKcs is more frequently altered in the basal-like subtype of breast cancer compared to other subtypes, which is associated with a higher fold change increase compared to the normal breast than other molecular subtypes. However, the relatively weak correlation between mRNA and protein expression in a subset of TCGA samples suggests that expression at the mRNA level may not be a good marker for the level and/or activity of the DNA-PKcs protein, which is known to be regulated by multiple post-translational modifications.

**Validation of the interaction between ELF5 and DNA-PKcs using co-immunoprecipitation**

In order to validate the interaction between ELF5 and DNA-PKcs, co-immunoprecipitation (co-IP) using cross-linked MCF7-ELF5-Isoform2-V5 nuclear lysates (prepared using the RIME protocol) was performed. Approximately 10% of the antibody-bound magnetic beads used for the immunoprecipitation (IP) of ELF5 and V5 (or IgG control) were reserved after the final wash step, with the remaining 90% resuspended in AMBIC solution in preparation for RIME. The reserved beads were resuspended in loading buffer with 2x reducing agent and heated to elute the immunoprecipitated proteins. These samples were then run on a 4-12% Bis-Tris gel, along with input samples (nuclear lysate prior to IP) and the supernatants from each IP and the first and third (final) bead washes. The co-IP results for three biological replicates are shown in Figure 5.5. The first replicate was also used for RIME (experiment 5), while additional replicates 1 and 2 were prepared using the RIME protocol but were not used for any ELF5-V5 RIME mass spectrometry experiments. In all three replicates, a DNA-PKcs band was seen at longer exposures in the ELF5-V5 IP (lane 2) but not in the IgG control IP (lane 3). There was also a large amount of DNA-PKcs protein seen in the supernatants from the IPs and first washes, suggesting that only a small fraction of the total cellular DNA-PKcs interacts with ELF5. However, this band was no longer visible in the final wash, indicating that the DNA-PKcs identified in the ELF5-V5 IPs was not a contaminant. Co-immunoprecipitation of cross-linked MCF7 cells therefore confirmed the interaction between ELF5-V5 and DNA-PKcs that was discovered by ELF5-V5 RIME.



**Figure 5.5: ELF5-V5 and DNA-PKcs co-immunoprecipitate in MCF7 cells**  
(previous page)

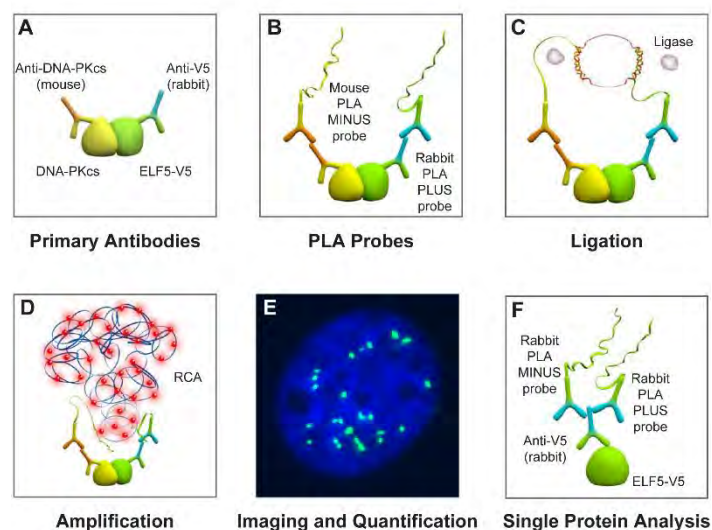
Samples were prepared using the RIME protocol and immunoprecipitated with a combination of ELF5 and V5 antibodies or IgG control. Blots for V5 and DNA-PKcs are shown. Lane 1 is the input or total lysate (In), lane 2 is the ELF5-V5 immunoprecipitation (IP) and lane 3 is the IgG control immunoprecipitation (IgG). Lanes 4 and 7 are supernatants from the immunoprecipitations (Sup, representing unbound protein), while lanes 5-6 and 8-9 are supernatants from the first (W1) and third (W3) bead washes (indicating no residual unbound protein after the final third wash). RIME 5 (top) is also replicate 5 of the ELF5-V5 RIME experiments, while additional replicates 1 and 2 do not form part of the ELF5-V5 RIME dataset.

**Validation of the interaction between ELF5 and DNA-PKcs using Proximity Ligation Assays**

The Duolink Proximity Ligation Assay (PLA) was also used to validate the interaction between ELF5-V5 and DNA-PKcs. PLA, summarised in Figure 5.6, is an immunofluorescence-based technique that results in generation of a fluorescent signal when two candidate proteins interact. PLA can also be adapted for high-sensitivity detection of a single protein (using one antibody with two same-species secondary PLA probes), as shown in Figure 5.6F.

The antibodies used were tested by single-antibody PLAs, as shown in Figures 5.7 (V5 antibody), 5.8A (DNA-PKcs antibody from Cell Signaling Technology, CST) and 5.8B (DNA-PKcs antibody from Thermo-Fisher, TF). The number of signals in the V5 single-antibody PLA in MCF7-ELF5-Isoform2-V5 cells was greatly increased when cells were treated with doxycycline (Figure 5.7, column i) compared to vehicle (column ii). The signal number in the MCF7-pHUSH-Empty line was also low (column iii), comparable to that seen in the rabbit IgG negative control (column iv). The V5 signals were predominantly nuclear but varied in quantity between cells, most likely due to variable expression in the pooled cell population as seen in previous ELF5-V5 immunofluorescence studies (Chapter 3). Similarly, the two DNA-PKcs antibodies (Figures 5.8A and 5.8B) produced strong nuclear signals in the single-antibody PLAs, consistent with the known subcellular localisation of this protein (columns i and ii). The signal quantity varied between cells but was not obviously affected by ELF5 overexpression or by the addition of doxycycline (empty vector cells, images not shown). However, quantification could not be reliably performed in the single-antibody PLAs as the large signal numbers resulted in a significant amount of coalescence.





**Figure 5.6: The immunofluorescence-based Proximity Ligation Assay (PLA) identifies interacting proteins**

*Figure adapted from User Guide Duolink In Situ - Fluorescence (Sigma-Aldrich, 2013)*

(A) Fixed cells are incubated with primary antibodies targeting candidate interacting proteins. The primary antibodies must be from different species. The example of V5 (rabbit antibody) and DNA-PKcs (mouse antibody) is shown. (B) Secondary antibodies conjugated to distinct oligonucleotides (PLA minus probe and PLA plus probe) are applied. In this example, the mouse minus probe recognises the mouse anti-DNA-PKcs antibody and the rabbit plus probe recognises the rabbit V5 antibody. (C) Ligation solution (containing minus and plus oligonucleotides and ligase) is added to the cells and incubated. The minus and plus oligonucleotides (red) hybridise to the respective probes and, if they are in close proximity ( $<40\text{nm}$ ), will join to form a closed circle. (D) Amplification solution (containing nucleotides, fluorescent oligonucleotides and polymerase) is added. The oligonucleotide arm of one of the probes acts as a primer for a rolling-circle amplification (RCA) reaction using the ligated circle as a template. The fluorescently labelled oligonucleotides hybridise to the repeated sequence RCA product. Extensive washes are performed between each of the steps A-D. (E) The fluorescently labelled product is easily visible as a fluorescent spot (PLA signal) using microscopy. Signals can be quantified using image analysis software. (F) (F) PLA may also be used for visualising a single protein with high sensitivity. Fixed cells are incubated with a single primary antibody (in this example, a rabbit antibody targeting V5) followed by incubation with rabbit PLA MINUS and rabbit PLUS probes. Both MINUS and PLUS probes will bind to the rabbit V5 antibody, resulting in generation of a fluorescent signal.

Figure 5.7: ELF5-V5 single-antibody PLA optimisation

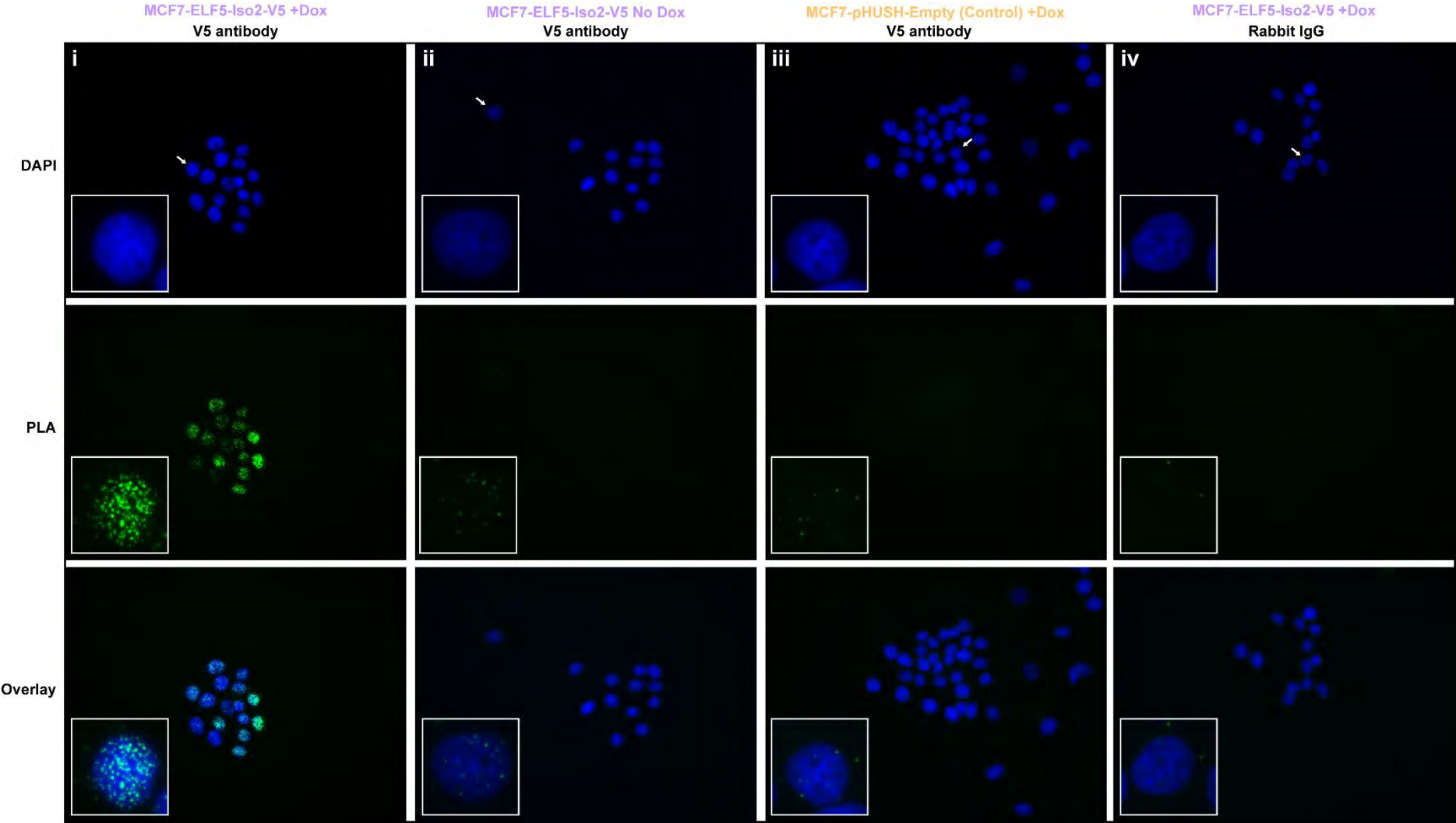


Figure 5.8A: DNA-PKcs CST single-antibody PLA optimisation

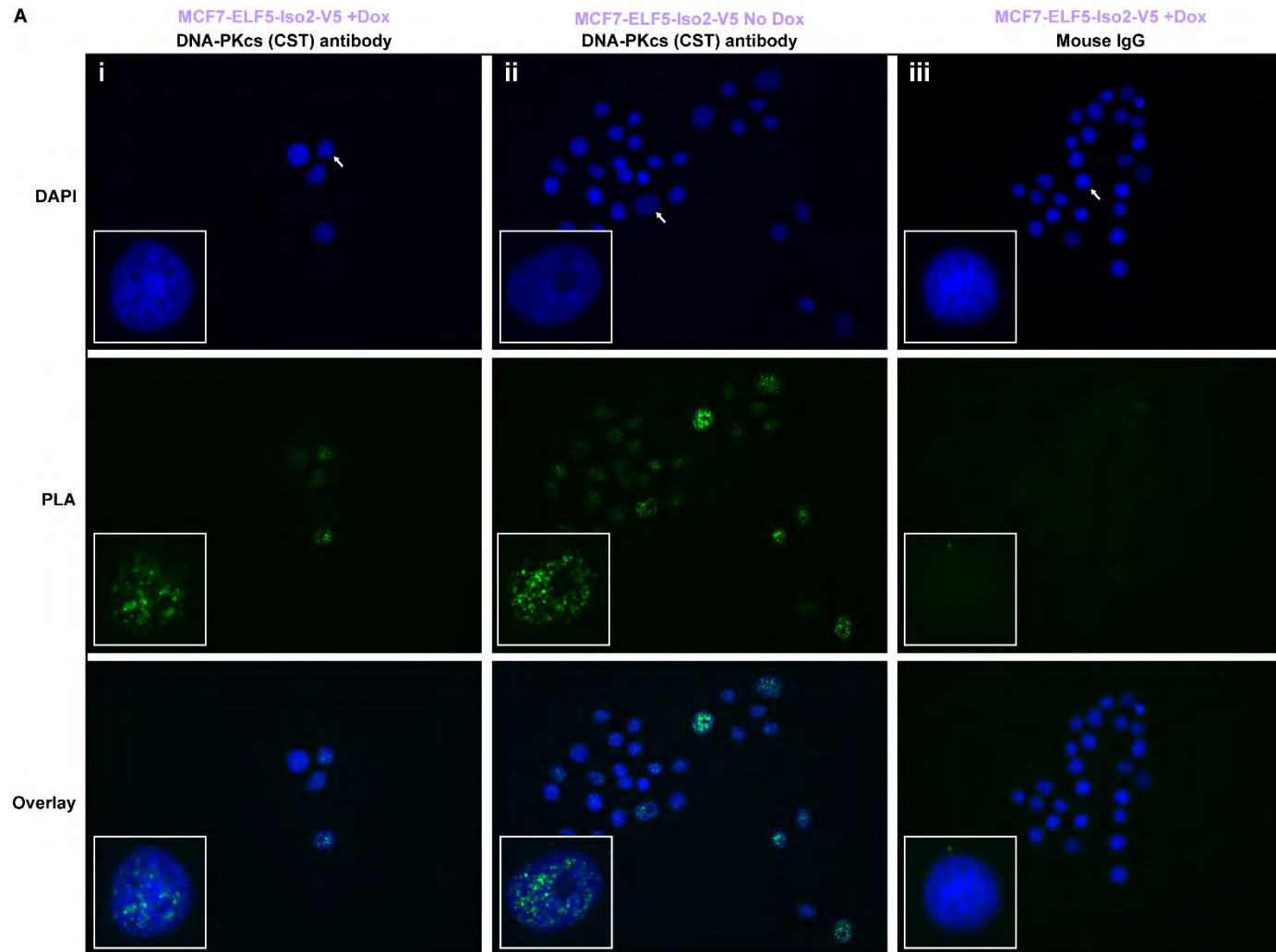
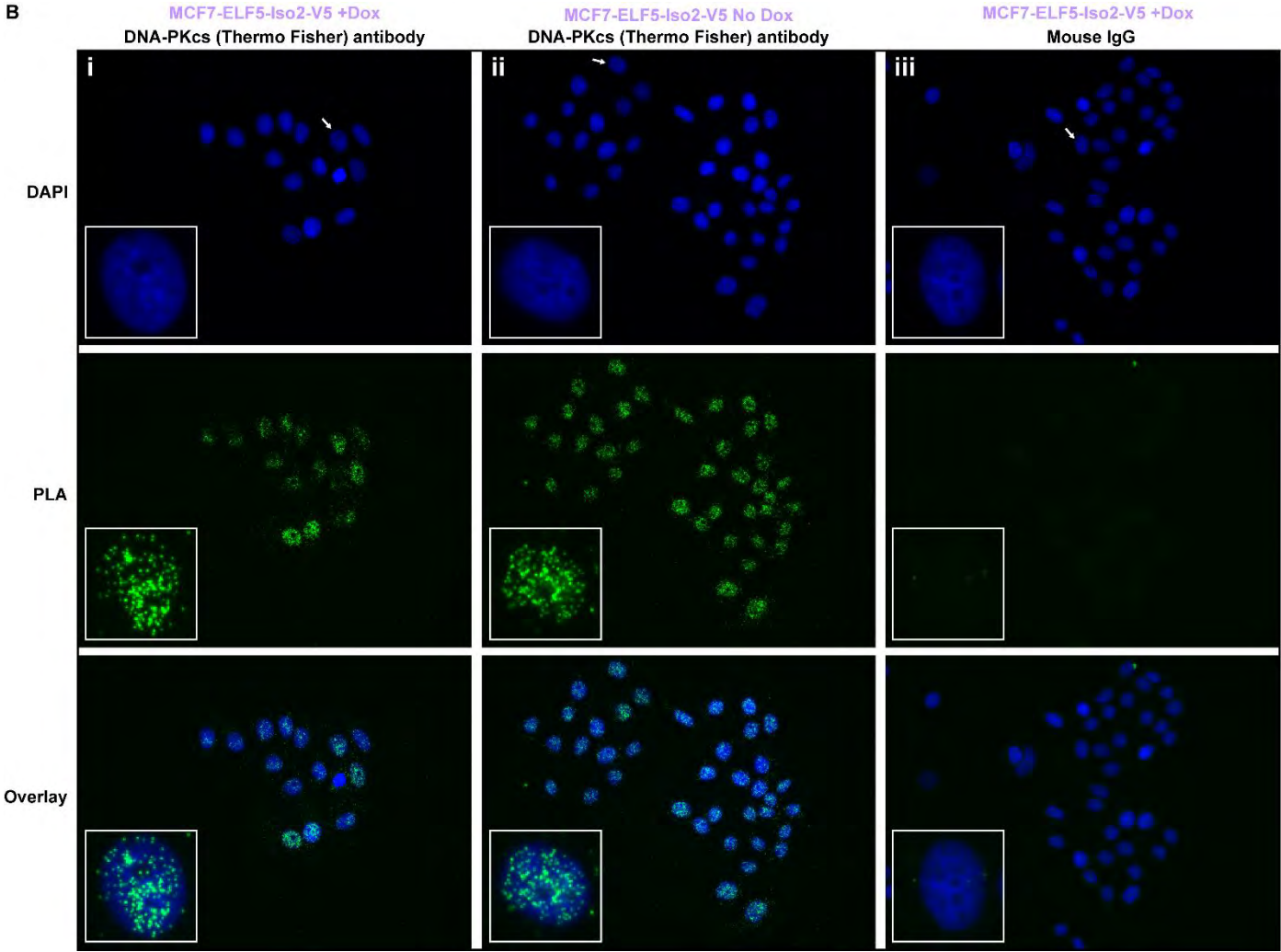


Figure 5.8B: DNA-PKcs Thermo single-antibody PLA optimisation



### **Figure 5.7: ELF5-V5 single-antibody PLA optimisation**

Single-antibody proximity ligation assays (PLAs) using a rabbit V5 antibody, with images arranged in columns: (i) MCF7-ELF5-Isoform2-V5 cells treated with doxycycline, (ii) MCF7-ELF5-Isoform2-V5 cells treated with vehicle (negative control), and (iii) MCF7-pHUSH-Empty cells treated with doxycycline (negative control). Single-antibody PLA using rabbit IgG (iv) was performed as an additional negative control. Top row is cell nuclei stained with DAPI, middle row is PLA signals and bottom row is the overlay image. Inset shows a representative enlarged nucleus, indicated by the white arrow in the DAPI image.

### **Figure 5.8: DNA-PKcs CST and DNA-PKcs Thermo single-antibody PLA optimisation**

Single-antibody proximity ligation assays (PLAs) using a mouse DNA-PK antibody from Cell Signaling Technology (CST) (panel A) or a mouse DNA-PK antibody from Thermo Fisher (TF) (Panel B), with images arranged in columns: (i) MCF7-ELF5-Isoform2-V5 cells treated with doxycycline and (ii) MCF7-ELF5-Isoform2-V5 cells treated with vehicle. Single-antibody PLA using mouse IgG (iii) is a negative control. Top row is cell nuclei stained with DAPI, middle row is PLA signals and bottom row is the overlay image. Inset shows a representative enlarged nucleus, indicated by the white arrow in the DAPI image.

To examine the interaction between ELF5-V5 and DNA-PKcs, PLAs were performed in MCF7 cell lines using a combination of V5 and DNA-PKcs (either CST or TF) antibodies. Example PLA images of MCF7-ELF5-V5 cells, either in the presence or absence of doxycycline, are shown in Figures 5.9A (V5 and DNA-PKcs CST) and 5.9B (V5 and DNA-PKcs TF). In addition to the images shown, numerous negative PLA controls were performed, including MCF7-pHUSH-empty vector cells (-/+ doxycycline), single IgG substitutions in doxycycline-treated MCF7-ELF5-V5 cells (for example, a combination of rabbit IgG and DNA-PKcs antibody), double IgG substitutions and a no primary antibody control. The representative images in Figures 5.9A and 5.9B show the presence of multiple signals in the MCF7-ELF5-V5 cells treated with doxycycline, indicating interactions between ELF5-V5 and DNA-PKcs. This was evident with both combinations of antibodies, although the DNA-PKcs Thermo-Fisher antibody combination produced a lower average signal number in doxycycline-treated cells and a higher level of background in the negative controls compared to the DNA-PKcs CST antibody combination.

Figure 5.9A: Proximity ligation assays corroborate the ELF5-DNA-PKcs nuclear interaction (V5 + DNA-PKcs CST combination)

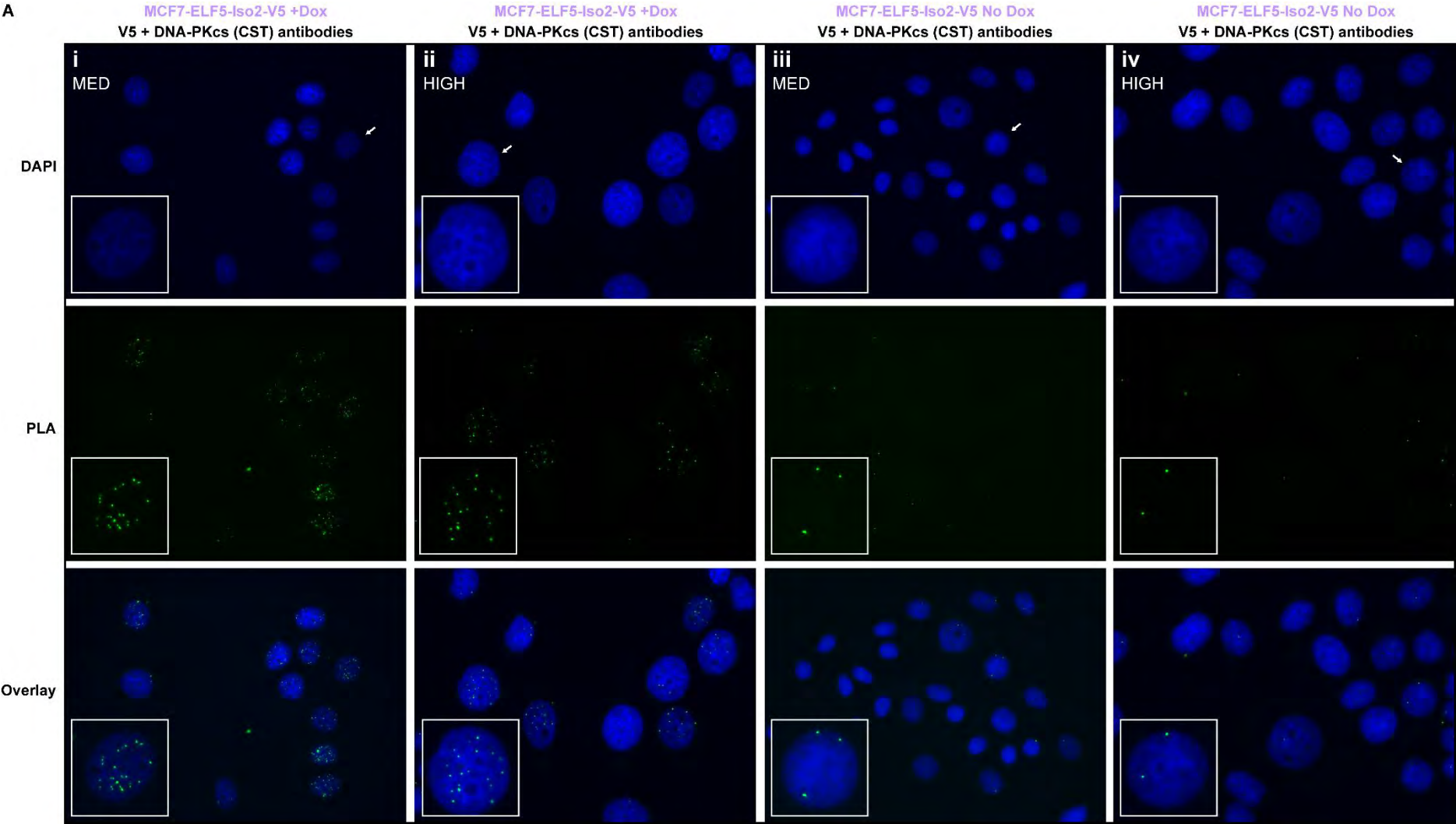
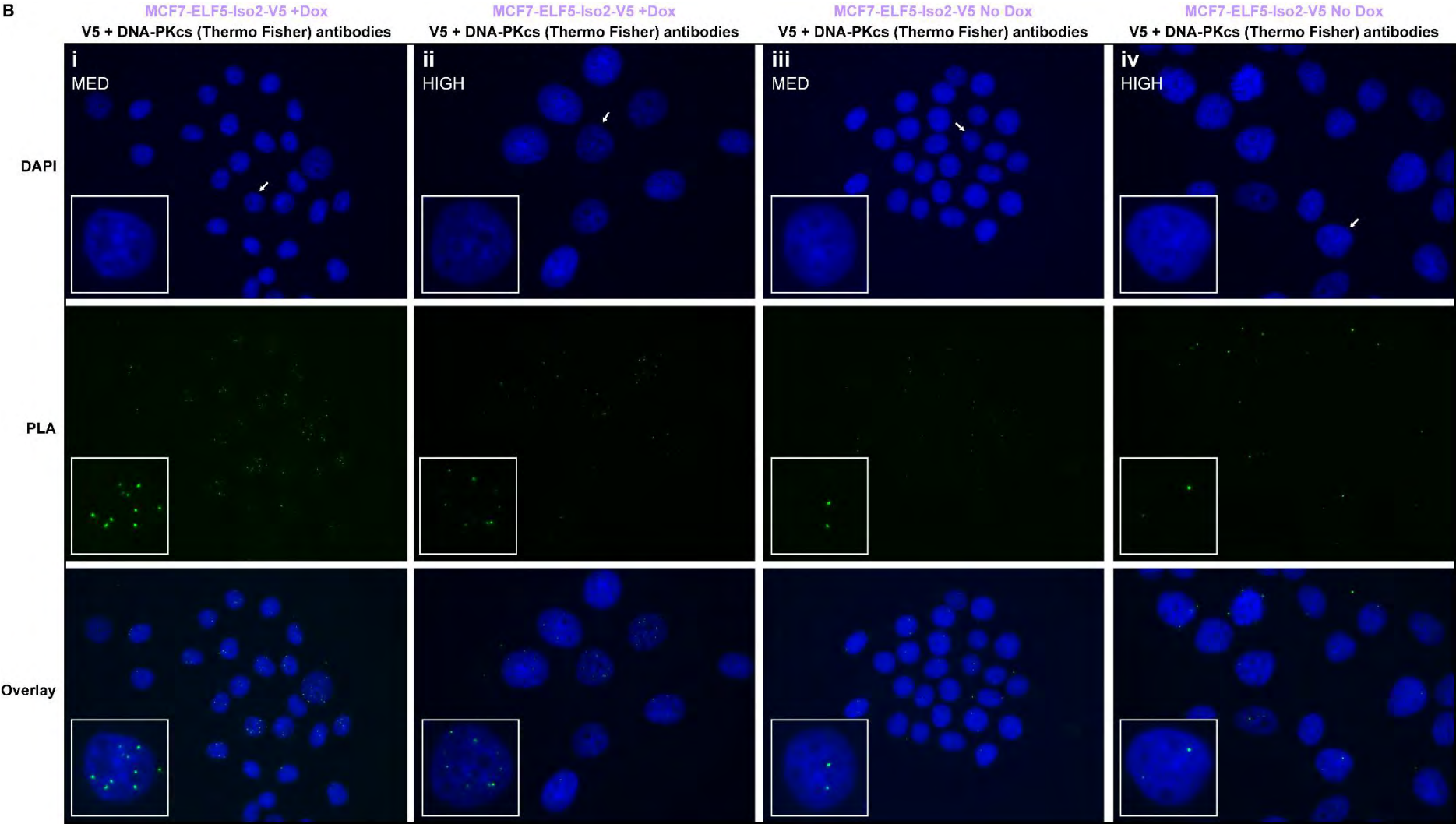




Figure 5.9B: Proximity ligation assays corroborate the ELF5-DNA-PKcs nuclear interaction (V5 + DNA-PKcs Thermo Fisher combination)

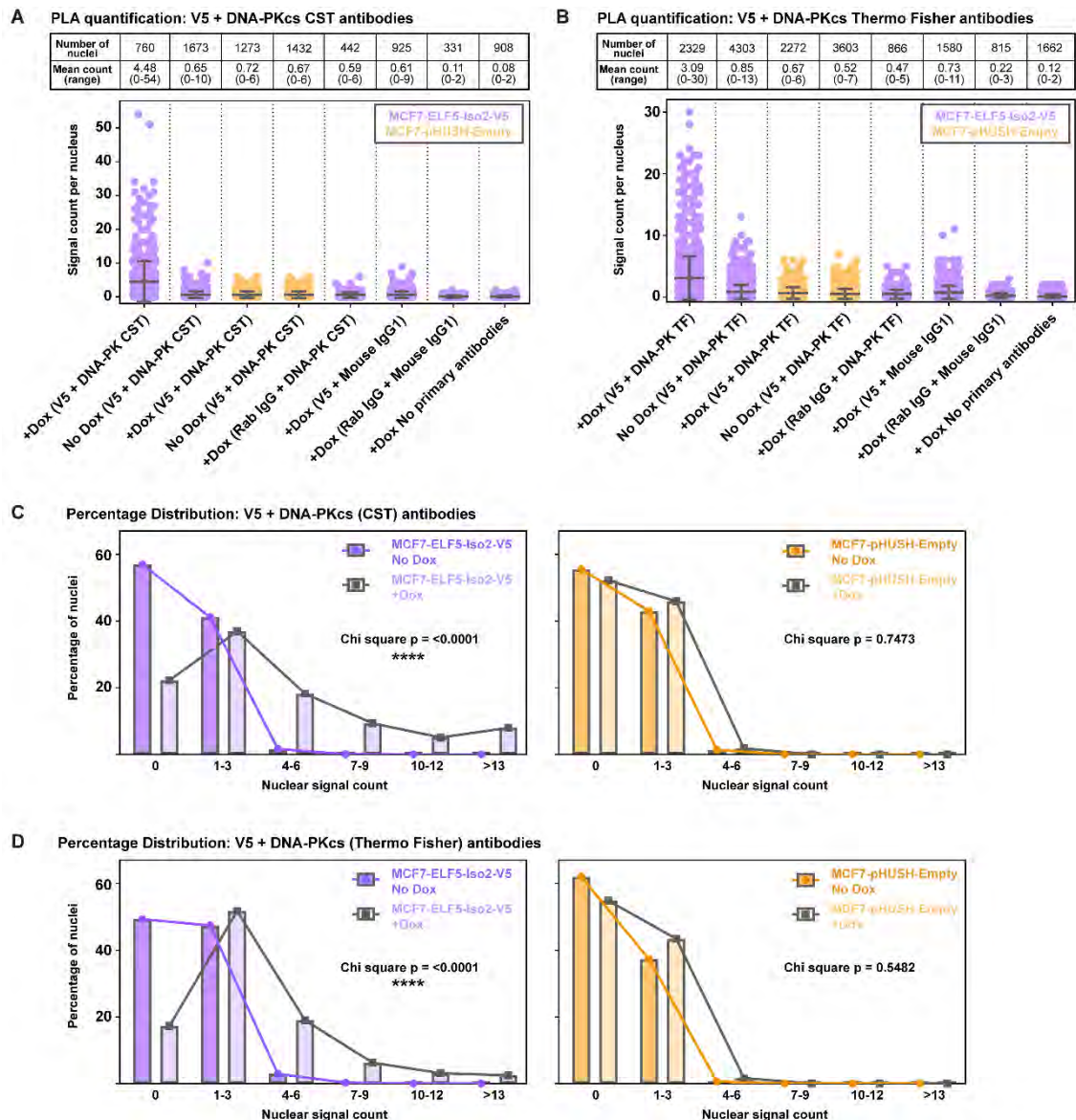


**Figure 5.9: Proximity ligation assays corroborate the ELF5-DNA-PKcs nuclear interaction**

Double-antibody proximity ligation assays using the rabbit V5 antibody and the mouse DNA-PK antibody from Cell Signaling Technology (CST) (panel A) or Thermo Fisher (TF) (panel B). Images, arranged in columns, show: (i-ii) MCF7-ELF5-Isoform2-V5 cells treated with doxycycline and (iii-iv) MCF7-ELF5-Isoform2-V5 cells treated with vehicle, at medium (MED) or high (HIGH) magnification. Top row is cell nuclei stained with DAPI, middle row is PLA signals and bottom row is the overlay image. Inset shows a representative enlarged nucleus, indicated by the white arrow in the DAPI image.

Quantification of the double-antibody PLAs, including all negative controls, is shown in Figure 5.10. Only nuclear signals were considered due to the predominantly nuclear localisation of both ELF5-V5 and DNA-PKcs. The average signal number in MCF7-ELF5-V5 cells treated with doxycycline was increased compared to all other conditions with both antibody combinations (Figures 5.10A and 5.10B). The distribution of nuclear PLA signals per cell in MCF7-ELF5-V5 (left) and MCF7-pHUSH-empty vector cells (right) was also analysed (Figures 5.10C and 5.10D). A chi-square test demonstrated that there was a significant change in the distribution of nuclear signals with both antibody combinations in the doxycycline-treated MCF7-ELF5-V5 (but not empty-vector) cells. In the V5 and DNA-PKcs CST double-antibody PLA (Figure 5.10C), 57% of vehicle-treated nuclei had no signals, compared to only 22% of doxycycline-treated nuclei. The majority of doxycycline-treated nuclei had 1-3 (37%) or 4-6 (18%) signals. Approximately 23% of doxycycline-treated cells had more than 7 signals per nucleus, compared to only 0.18% of vehicle-treated cells. The MCF7-pHUSH-Empty cells (-/+ doxycycline) show an even more dramatic decline in distribution, with 98% of all nuclei having 3 or less signals and no nuclei having more than 6 signals. Similar results were seen in the V5 and DNA-PKcs TF double-antibody PLA (Figure 5.10D), with 49% of vehicle-treated cells and only 17% of doxycycline-treated cells having no signals. The majority of doxycycline-treated cells had 1-3 (52%) or 4-6 (19%) signals, with approximately 12% of cells having more than 7 signals per nucleus (compared to only 0.33% of vehicle-treated cells). The increase in the average PLA signal number per nucleus in doxycycline-treated cells, combined with the significant change in signal distribution, indicate a clear interaction between ELF5-V5 and DNA-PKcs in MCF7 cells. PLA therefore provided an independent and quantifiable method for validation of ELF-V5 RIME results.





**Figure 5.10: Quantification of ELF5-DNA-PKcs PLA signals in MCF7 cell lines**

(A-B) Double-antibody PLA signal counts for MCF7-ELF5-Isoform2-V5 (purple) and MCF7-pHUSH-Empty (yellow) cell lines, using the rabbit V5 antibody and mouse DNA-PK antibody from Cell Signaling Technology (A) or Thermo Fisher (B). Plotted values represent the signal counts for individual nuclei and error bars the mean with standard deviation. X-axis labels indicate doxycycline (+Dox) or vehicle (No Dox) treatment and the combination of antibodies used for the PLA in parentheses. IgG substitutions were made for one or both antibodies as negative controls. The tables above the graphs indicate the number of nuclei counted for a given condition (top row) and the average number of signals per nucleus (bottom row), with the range of counts per nucleus indicated in parentheses. (C-D) The percentage of nuclei with a signal count in the indicated range using the rabbit V5 antibody and DNA-PK antibody from Cell Signaling Technology (C) or Thermo Fisher (D). Cells were treated with doxycycline (light shading) or vehicle (dark shading), with MCF7-ELF5-Isoform2-V5 cells on the left (purple) and MCF7-pHUSH-Empty cells on the right (yellow). A chi-square test was used to calculate p-values.

## Phosphorylation of ELF5

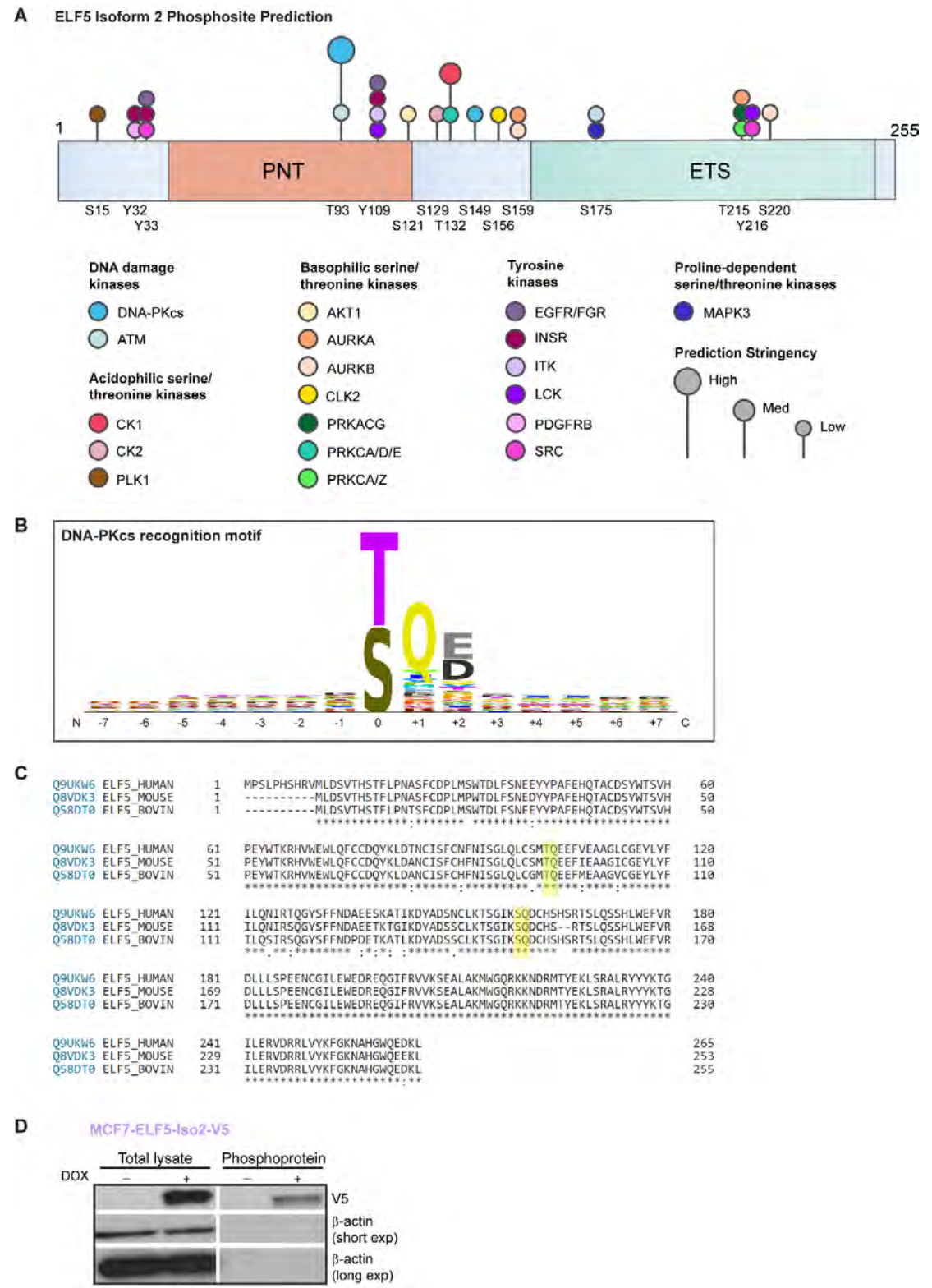
Most functions of DNA-PKcs, including transcriptional regulation, require the intact activity of the kinase domain. DNA-PKcs is known to interact with and phosphorylate many transcription factors and co-factors. The next step was therefore to determine if ELF5 was in fact phosphorylated and, if so, whether DNA-PKcs was one of the kinases involved.

A motif analysis of the human ELF5 protein sequence (Isoforms 1 and 2) was performed to identify potential phosphorylation sites (Figure 5.11A) (Obenauer *et al.*, 2003). The highest stringency settings (top 0.2% of motif matches within the vertebrate database) identified a single candidate phosphorylation site at threonine 93 (T93) of ELF5 Isoform 2, predicted to be catalysed by DNA-PKcs. T93 lies within the Pointed domain of ELF5 and is also associated with increased surface accessibility as predicted by protein sequence. The optimal motif for DNA-PKcs is shown in Figure 5.11B and consists of a threonine or serine at position 0, preferably followed by a glutamine (Q) at +1 and a glutamic acid (E) or aspartic acid (D) at +2. As can be seen in Figure 5.11C, T93 (or T103 as shown here in ELF5 Isoform 1) lies within the sequence context TQE, which is conserved across human, mouse and cow. The T93 site received a Scansite score of 0.361 (where 0.000 represents the optimal motif match), placing it in the top 0.056% of matches in the vertebrate database with a z-score (standard deviations away from the mean) of -4.47.

At low stringency settings (top 5% of motif matches within the vertebrate database), 14 additional potential phosphorylation sites were identified, potentially catalysed by as many as 23 different kinases (Figure 5.11A). Within this set, one additional DNA-PKcs site was predicted at serine 149 (S149), in the region between the Pointed and ETS domains. S149 (S159 as shown in Figure 5.11C) lies within the sequence SQD, which is also conserved across the three species shown. This site received a Scansite score of 0.564, placing it in the top 1.97% of matches with a z-score of -2.73.

Phosphoprotein purification, utilising specialised columns containing a resin that binds phosphorylated proteins, was performed using lysates from MCF7-ELF5-V5 cells +/- doxycycline (Figure 5.11D). ELF5-V5 was recovered in the phosphoprotein fraction in doxycycline-treated cells, demonstrating for the first time that ELF5 is phosphorylated. In contrast, no  $\beta$ -actin was seen in the phosphoprotein fraction despite being abundant in the total lysate. Although phosphorylation sites have been identified in  $\beta$ -actin (for example, in the study by Sharma *et al.*), it appeared in this context to be non-

phosphorylated and therefore functioned as an effective negative control for contamination of the phosphoprotein fraction by abundant cellular proteins. While phosphoprotein purification demonstrated that ELF5-V5 was phosphorylated in MCF7-ELF5-V5 cells, no conclusions about the specific site of phosphorylation or the kinases responsible could be drawn from this approach.



### Figure 5.11: ELF5 is a phosphoprotein *in vivo*

(previous page)

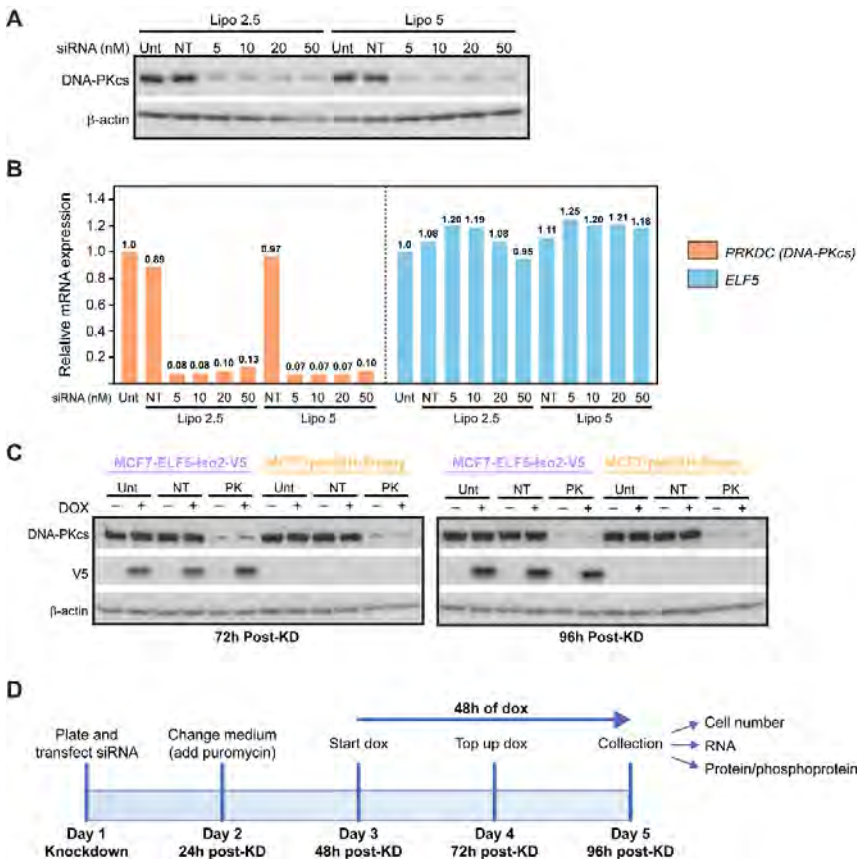
(A) Prediction of phosphorylation sites in the ELF5 protein using Scansite 3 software (<http://scansite3.mit.edu>) (Obenauer *et al.*, 2003). The Pointed (PNT) and ETS domains of ELF5 are indicated. The predicted phosphosites are represented by circles and labelled with the amino acid residue and number. Predictions were made using high stringency (top 0.2% of motif matches within the vertebrate database), medium stringency (top 1%) and low stringency (top 5%) settings, indicated by large/medium/small circles respectively. The kinases predicted to phosphorylate these sites are colour-coded as shown. (B) Recognition motif for DNA-PKcs from Scansite, consisting of a serine or threonine at position 0, preferably followed by a glutamine ("Q") at +1 and a glutamic acid ("E") or aspartic acid ("D") at +2. (C) Conservation of ELF5 protein sequence in human (Isoform 1 numbering shown), mouse and cow. Asterisk on bottom row indicates conserved residue. The two predicted DNA-PKcs phosphorylation sites are highlighted in yellow. (D) Phosphoprotein purification of MCF7-ELF5-Isoform2-V5 cells treated with doxycycline or vehicle. Total lysates are loaded on the left side of the gel and the phosphoprotein fractions are loaded on the right. Short and long exposures (exp) are shown for  $\beta$ -actin. Amino acid abbreviations: D, aspartic acid; E, glutamic acid; Q, glutamine; S, serine; T, threonine; Y, tyrosine. Kinases: AKT1, AKT serine/threonine kinase 1; ATM, ataxia telangiectasia mutated serine/threonine kinase; AURKA, aurora kinase A; AURKB, aurora kinase B; CK1, casein kinase 1; CK2, casein kinase 2; CLK2, CDC like kinase 2; DNA-PKcs, DNA-dependent protein kinase catalytic sub-unit; EGFR, epidermal growth factor receptor; FGR, FGR proto-oncogene, Src family tyrosine kinase; INSR, insulin receptor; ITK, Interleukin-2-inducible T-cell kinase; LCK, LCK proto-oncogene, Src family tyrosine kinase; MAPK3, mitogen-activated protein kinase 3 (also known as ERK1); PDGFRB, platelet derived growth factor receptor beta; PLK1, polo-like kinase 1; PRKACG, protein kinase cAMP-activated catalytic sub-unit gamma; PRKCA, protein kinase C alpha; PRKCD, protein kinase C delta; PRKCE, protein kinase C epsilon; PRKCZ, protein kinase C zeta; SRC, SRC proto-oncogene, non-receptor tyrosine kinase.

### Optimisation of siRNA-mediated knockdown of DNA-PKcs in breast cancer cell lines

In order to investigate the functional relationship between these two proteins, knockdown of DNA-PKcs was optimised in the breast cancer cell line MCF7-ELF5-V5. The initial optimisation experiment indicated that a reasonable knockdown was achieved after 48 hours at both the protein (Figure 5.12A) and mRNA (Figure 5.12B) level with an siRNA concentration as low as 5nM. The knockdown level of 7-8% of untransfected cells by mRNA was similar using either 2.5ul or 5ul of transfection reagent per well (6-well plate). There was also a small increase (up to 25%) noted in ELF5 mRNA expression with DNA-PKcs knockdown.

A further optimisation experiment (using 5nM siRNA and 5ul lipofectamine per well) established that doxycycline treatment commenced 24 hours after transfection did not significantly alter the level of knockdown. In addition, the knockdown could be effectively maintained for up to 96 hours post-transfection (Figure 5.12C). A previous report has suggested that doxycycline can reduce DNA-PKcs protein expression in breast cancer cell lines (Lamb *et al.*, 2015). However, no change in DNA-PKcs expression was seen in the empty vector control cell line treated with doxycycline (at a dose more than 200x lower than in the previous report).

Based on these experiments, a timeline for DNA-PKcs knockdown in combination with ELF5 overexpression was chosen (Figure 5.12D). In all subsequent knockdown experiments, cells were transfected on plating (day 0) using 5nM siRNA and 2.5ul lipofectamine per well (6-well plate, scaled as necessary according to plate size). The cells were given an additional day to recover after the siRNA transfection, with doxycycline commenced at 48 hours post-transfection to induce ELF5-V5 expression. Cells were collected at 96 hours post-transfection (a total of 48 hours of doxycycline treatment) for cell count, mRNA and protein / phosphoprotein analysis. Additional cell lines were tested using this method, with effective knockdown combined with ELF5 overexpression also achieved in T47D- and MDA-MB-231-ELF5-V5 cell lines.





**Figure 5.12: Efficient knockdown of DNA-PKcs can be achieved in breast cancer cells (previous page)**

(A) Western blots of MCF7-ELF5-Isoform2-V5 cells following transfection of siRNA against DNA-PKcs at increasing concentrations (5-50nM), using either 2.5uL or 5.0uL of Lipofectamine RNAiMAX (Lipo) per well of a 6-well tissue culture plate. Controls were untransfected (Unt) or transfected with non-targeting siRNA (NT). (B) Quantitative PCR for DNA-PKcs (left) and ELF5 (right) following transfection of siRNA against DNA-PKcs as above (single replicate). Relative gene expression values are shown for each sample, normalised to the untransfected control. (C) Western blots of MCF7-ELF5-Isoform2-V5 and MCF7-pHUSH-Empty cells following transfection of siRNA against DNA-PKcs at a concentration of 5nM using 5uL of Lipofectamine per well. Cells were treated with doxycycline (+) or vehicle (-) commencing 24 hours post-transfection and were collected at 72 hours (left) or 96 hours (right) post-transfection. (D) Timeline for all subsequent combined DNA-PKcs knockdown and ELF5 overexpression experiments. Cells were transfected on day 0 using 5nM siRNA and 2.5uL lipofectamine per well. On day 1, the culture medium was changed and puromycin was commenced to maintain doxycycline-induced ELF5 expression. Doxycycline treatment was started on day 2 and cells were collected for analysis of cell number, mRNA and protein on day 4. Lipo, Lipofectamine RNAiMAX; Unt, untransfected; NT, transfection with non-targeting siRNA; PK, transfection with siRNA against DNA-PKcs.

**ELF5-V5 phosphorylation in DNA-PKcs-knockdown cells**

One of the initial questions to be addressed was whether DNA-PKcs knockdown would affect ELF5-V5 phosphorylation. Doxycycline-treated MCF7-ELF5-V5 cells were transfected with siRNA targeting DNA-PKcs (or a non-targeting control siRNA, siNT) according to the above protocol (Figure 5.12D). The reduction in DNA-PKcs protein was confirmed by western blot of the total unfractionated lysate (data not shown). ELF5-V5 was present in the phosphoprotein fraction in all conditions tested (Figure 5.13A). The amount of ELF5-V5 in the phosphoprotein fraction was not reduced by DNA-PKcs knockdown and in fact was slightly increased, consistent with the small increase in total ELF5-V5 seen in the unfractionated lysate.

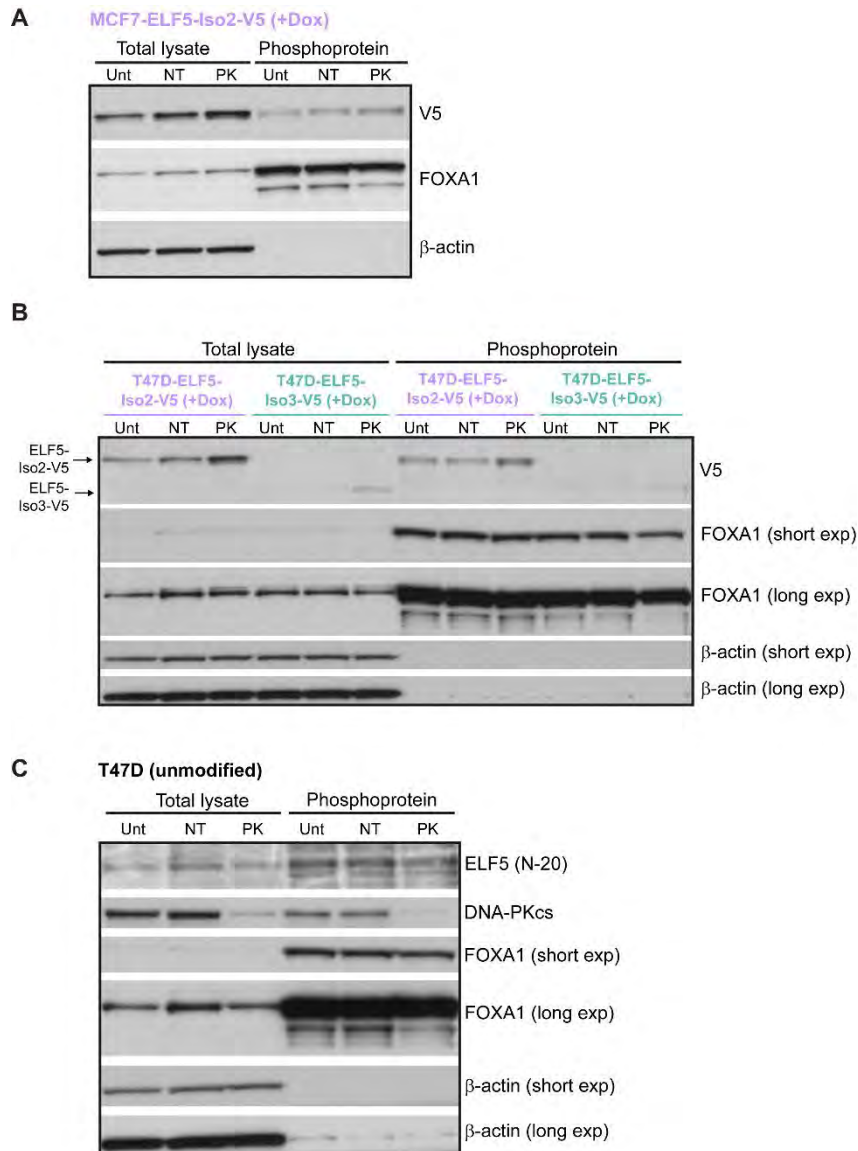
The experiment was also performed in T47D cell lines, including T47D-ELF5-Isoform2-V5 and T47D-ELF5-Isoform3-V5. ELF5 isoform 3 lacks the Pointed domain, which contains the more robustly predicted DNA-PKcs phosphorylation site. Again, ELF5-V5 was detected in all three conditions in both cell lines and was correlated with the level of ELF5-V5 expression in the total lysate, which was increased by DNA-PKcs knockdown (Figure 5.13B). The expression of ELF5-Isoform3-V5 was relatively low, however a faint band was visible in both the total lysate and phosphoprotein fractions.

This indicates that ELF5-V5 is phosphorylated at sites outside the Pointed domain, although it does not exclude the additional phosphorylation of Pointed domain residues.

In order to determine if phosphorylation also occurs when ELF5 is expressed at endogenous levels, phosphoprotein purification was performed in unmodified T47D cells. ELF5 could be detected in both the total lysate and the phosphoprotein fractions in T47D cells, confirming phosphorylation of endogenous ELF5 (Figure 5.13C).

In addition to ELF5-V5, the phosphoprotein fractions were immunoblotted for the transcription factor FOXA1. This was performed because high-stringency phosphosite analysis of the FOXA1 sequence identified two potential phosphorylation sites, including threonine 22 (T22) that was predicted to be catalysed by DNA-PKcs. The closely-related FOXA2 protein is known to be phosphorylated by AKT at T156 in response to insulin signalling (Wolfrum *et al.*, 2003) and by DNA-PKcs at serine 283 (Nock *et al.*, 2009), although corresponding residues are not present in FOXA1. Phosphoprotein purification confirmed that FOXA1 was phosphorylated in all MCF7 and T47D cell lines examined and was not obviously affected by either ELF5 overexpression or DNA-PKcs knockdown (Figures 5.13A-C).

The lack of effect of DNA-PKcs knockdown of ELF5 phosphorylation could be a result of several factors. Firstly, DNA-PKcs may not be the kinase responsible for ELF5 phosphorylation. It is also possible that DNA-PKcs is simply one of many kinases that act on ELF5, resulting in a small net effect of DNA-PKcs knockdown on total ELF5 phosphorylation levels. In fact, the lower stringency phosphosite prediction identified a total of 15 possible ELF5 phosphorylation sites, catalysed by up to 22 different kinases. While some of these are probably false-positives, it remains likely that more than one kinase phosphorylates ELF5 at various sites. In addition, the knockdown of DNA-PKcs was not complete. DNA-PKcs is a very highly expressed protein and although knockdown of more than 90% was achieved, there was a residual amount of protein remaining (as can be seen in Figure 5.13C). This small amount of protein may be sufficient for DNA-PKcs to phosphorylate some targets. In summary, although ELF5 was confirmed as a phosphoprotein in multiple breast cancer cell lines by these experiments, no further insight was gained into the site of ELF5 phosphorylation or whether DNA-PKcs phosphorylates ELF5.



**Figure 5.13: DNA-PKcs knockdown does not alter ELF5 phosphorylation**

(A) Phosphoprotein purification of doxycycline-treated MCF7-ELF5-Isoform2-V5 cells, with no siRNA transfection (Unt), non-targeting siRNA transfection (NT) or DNA-PKcs siRNA transfection (PK). Total lysates are loaded on the left side of the gel and the phosphoprotein fractions are loaded on the right. Western blots are for ELF5-V5, FOXA1 and β-actin.

(B) Phosphoprotein purification of doxycycline-treated T47D-ELF5-Isoform2-V5 (purple) and T47D-ELF5-Isoform3-V5 (green) cells with siRNA transfection as above. Total lysates are loaded on the left side of the gel and the phosphoprotein fractions are loaded on the right. Short and long exposures (exp) are shown for FOXA1 and β-actin. (C) Phosphoprotein purification for unmodified T47D cells with siRNA transfection as above, with total lysates on the left and phosphoprotein fractions on the right. Western blots are for endogenous ELF5 using the N-20 antibody, DNA-PKcs, FOXA1 (short and long exposures) and β-actin (short and long exposures). Unt, untransfected; NT, transfection with non-targeting siRNA; PK, transfection with siRNA against DNA-PKcs.

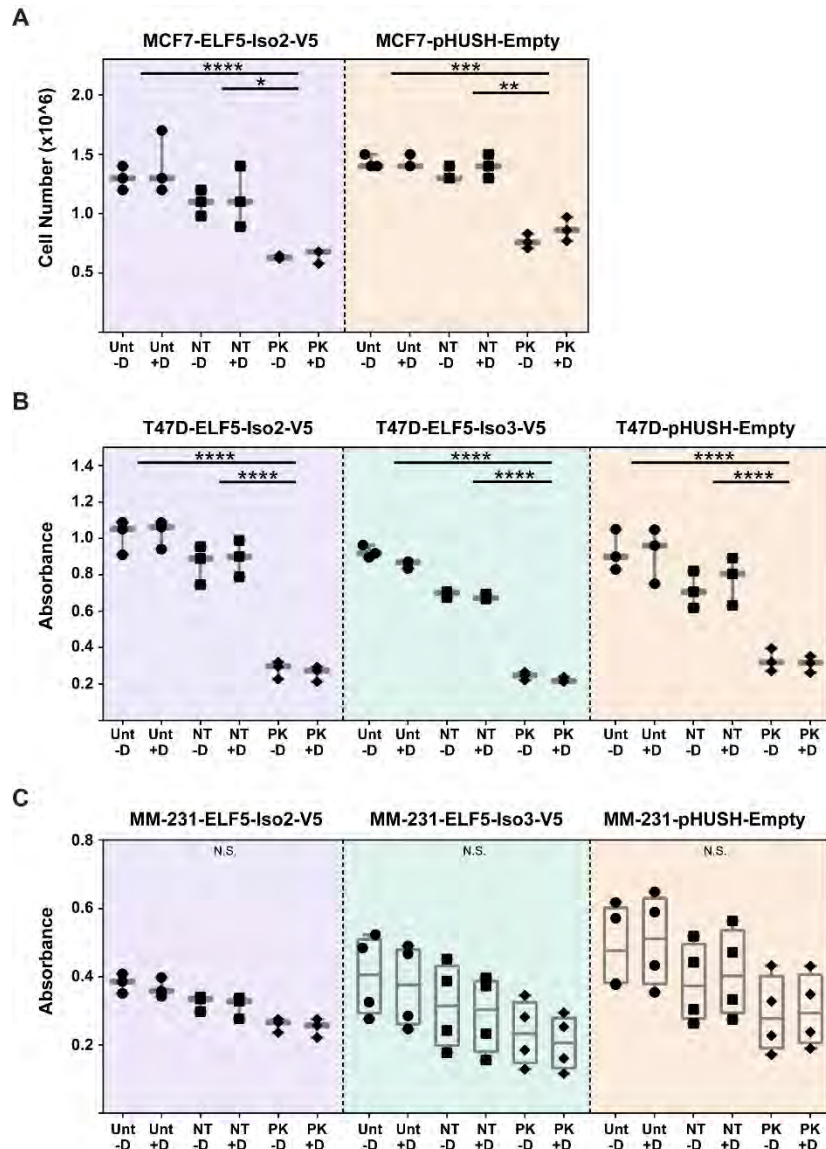


## **Phenotype of DNA-PKcs-knockdown cells**

Multiple cell lines were treated according to the previously described knockdown protocol. These included MCF7, T47D and MDA-MB-231 lines, with inducible expression of ELF5-Isoform2-V5 (all lines) or Isoform3-V5 (T47D and MDA-MB-231 lines only) and matched empty vector controls. In order to minimise selection effects, the lines chosen were pooled cell lines rather than clonal cell lines, although this is known to result in variable levels of ELF5 induction in individual cells. All experiments were conducted a minimum of three times.

An obvious phenotype began to appear in all lines with DNA-PKcs knockdown (including pHUSH-empty vector controls for ELF5) within 48 hours. This was most pronounced in the T47D lines and minimal in the MDA-MB-231 lines. In all lines, there was a clear decrease in the number of attached cells at day 4 (96 hours post-knockdown). This is quantified in Figures 5.14A-C for each cell line. The main differences in cell number were due to DNA-PKcs knockdown, with all MCF7 and T47D cell lines demonstrating a significant decrease in knockdown cells in both doxycycline- and vehicle-treated conditions. The MDA-MB-231-ELF5-Isoform2-V5 cell lines showed a downwards trend in the DNA-PKcs knockdown cells, however due to the smaller scale of difference as well as the large variation between replicates there were no statistically significant differences between conditions in these lines. There was a small decrease in cell number in the non-targeting control compared to the untransfected cells, however this did not reach statistical significance in any line. In addition, there was no significant difference between cell number in the matched doxycycline- and vehicle-treated samples in any cell line (for example, the untransfected cells -/+ doxycycline). This indicates that ELF5 overexpression is having a minimal effect on cell number.

Representative timecourse images (days 1-4 post-transfection) are shown in Figures 5.15A-C. The images shown are from the empty vector control cell lines, demonstrating that the phenotypic effects are primarily due to DNA-PKcs knockdown and not ELF5 overexpression. At 24 hours, the numbers of attached cells are similar, however from this point forwards the DNA-PKcs-knockdown cells accumulate at a slower rate. The MDA-MB-231 cells in particular also showed an increased number of floating cells at later timepoints.



**Figure 5.14: DNA-PKcs knockdown significantly reduces cell number in luminal breast cancer cell lines**

Cell number analysis on day 4 for MCF7- (A), T47D- (B) and MDA-MB-231- (C) ELF5-V5 cell lines treated with doxycycline (+D) or vehicle (-D). Cells were untransfected (Unt), transfected with non-targeting siRNA (NT) or transfected with siRNA against DNA-PKcs (PK). Values from 3-4 biological replicates are shown in a box-and-whisker plot, with the median represented by a horizontal line, the interquartile range represented by the box and the minimum to maximum range represented by the whiskers. Significant differences in counts between transfection conditions are shown according to the key below. There were no significant differences between matched doxycycline- and vehicle-treated cells in any cell line (for example, the untransfected cells +/- doxycycline). In the MCF7 lines, cell number was measured using an automated counter, while in the T47D and MDA-MB-231 lines cell number was estimated using a spectrophotometric absorbance assay. P-values (ANOVA):  $p < 0.0001$  (\*\*\*\*),  $p < 0.001$  (\*\*\*),  $p < 0.01$  (\*\*),  $p < 0.05$  (\*),  $p > 0.05$  (not significant, NS).

Figure 5.15A: DNA-PKcs knockdown alters cell phenotype over a 4-day timecourse (MCF7 cells)

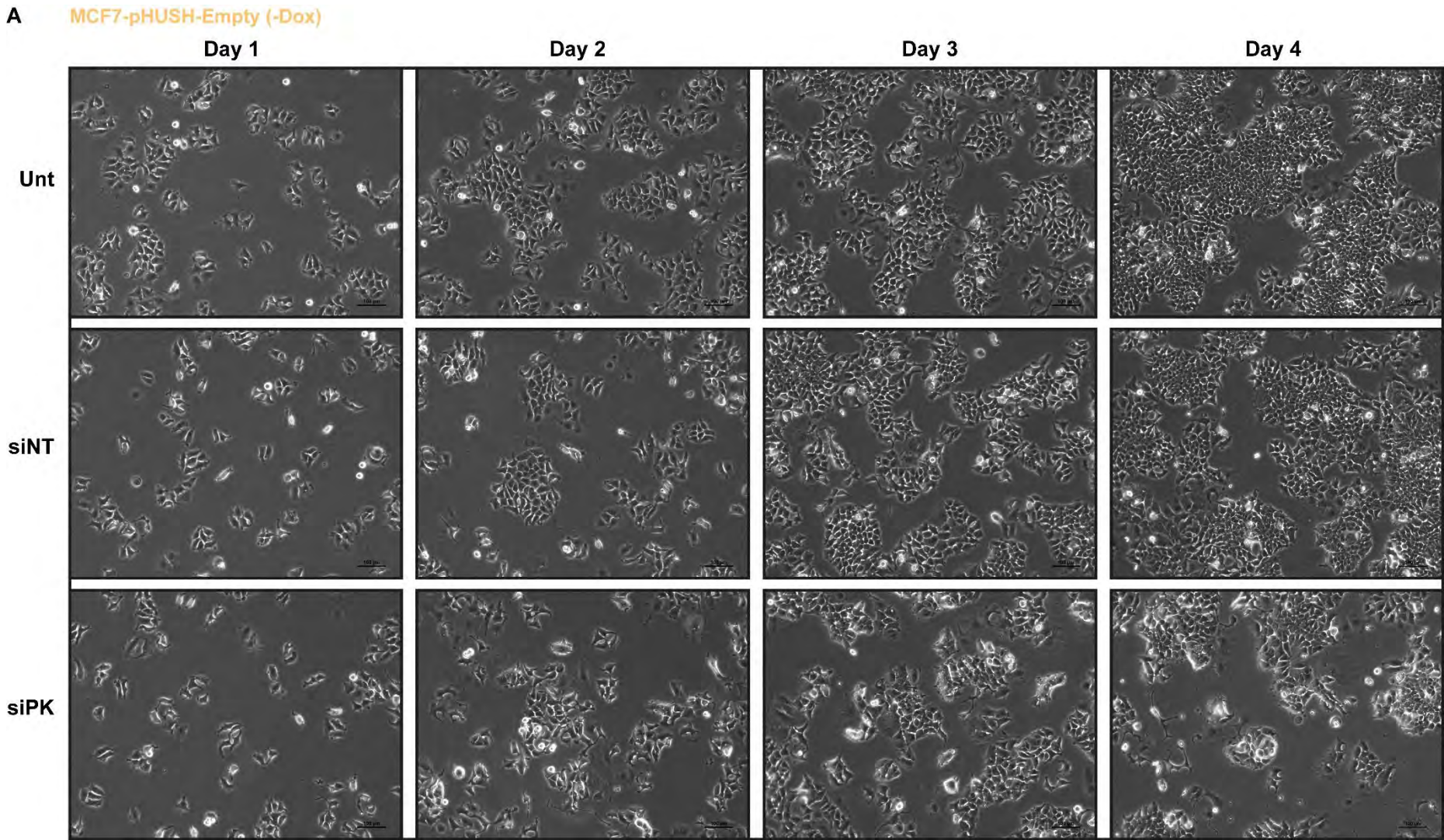




Figure 5.15B: DNA-PKcs knockdown alters cell phenotype over a 4-day timecourse (T47D cells)

B T47D-pHUSH-Empty (-Dox)

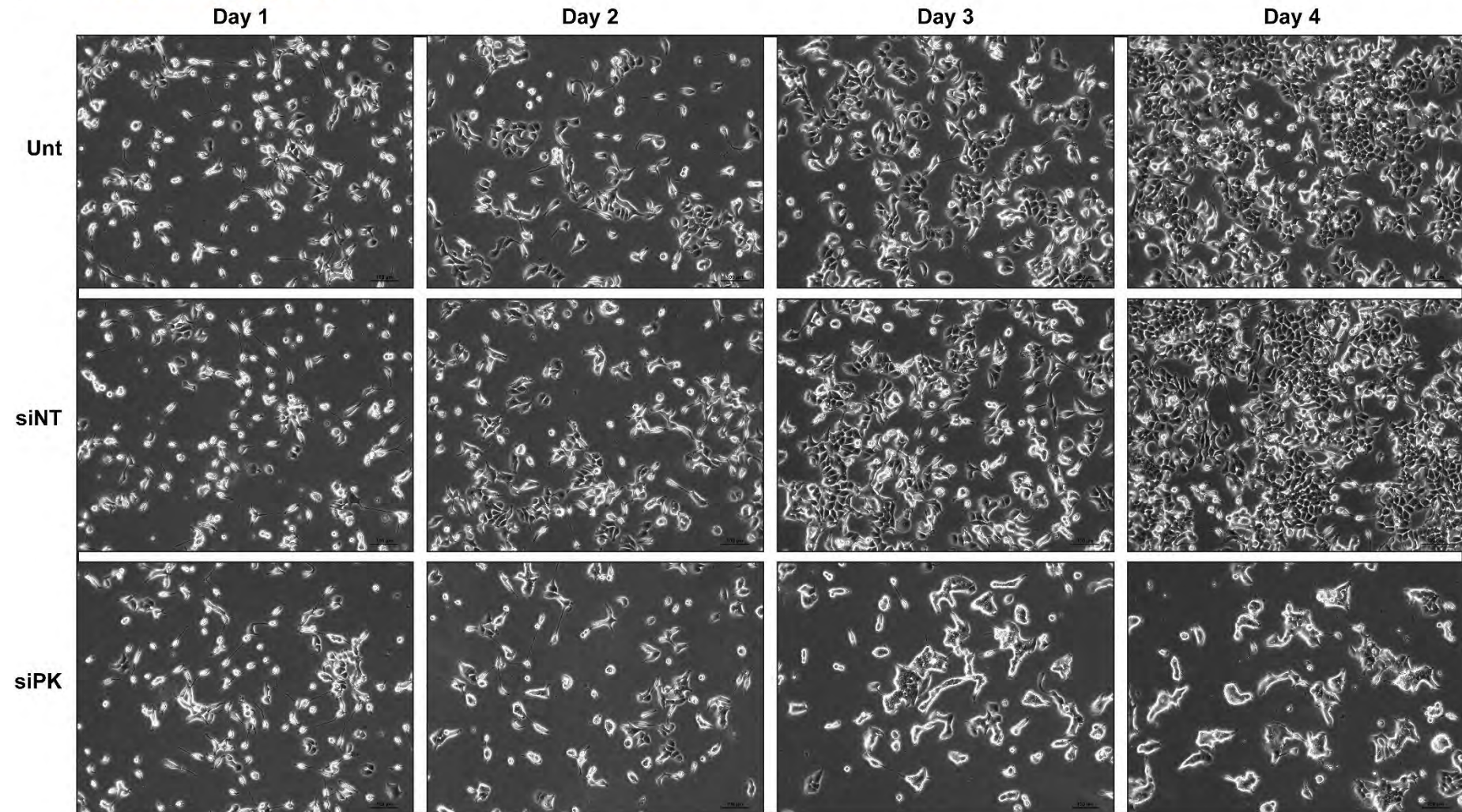
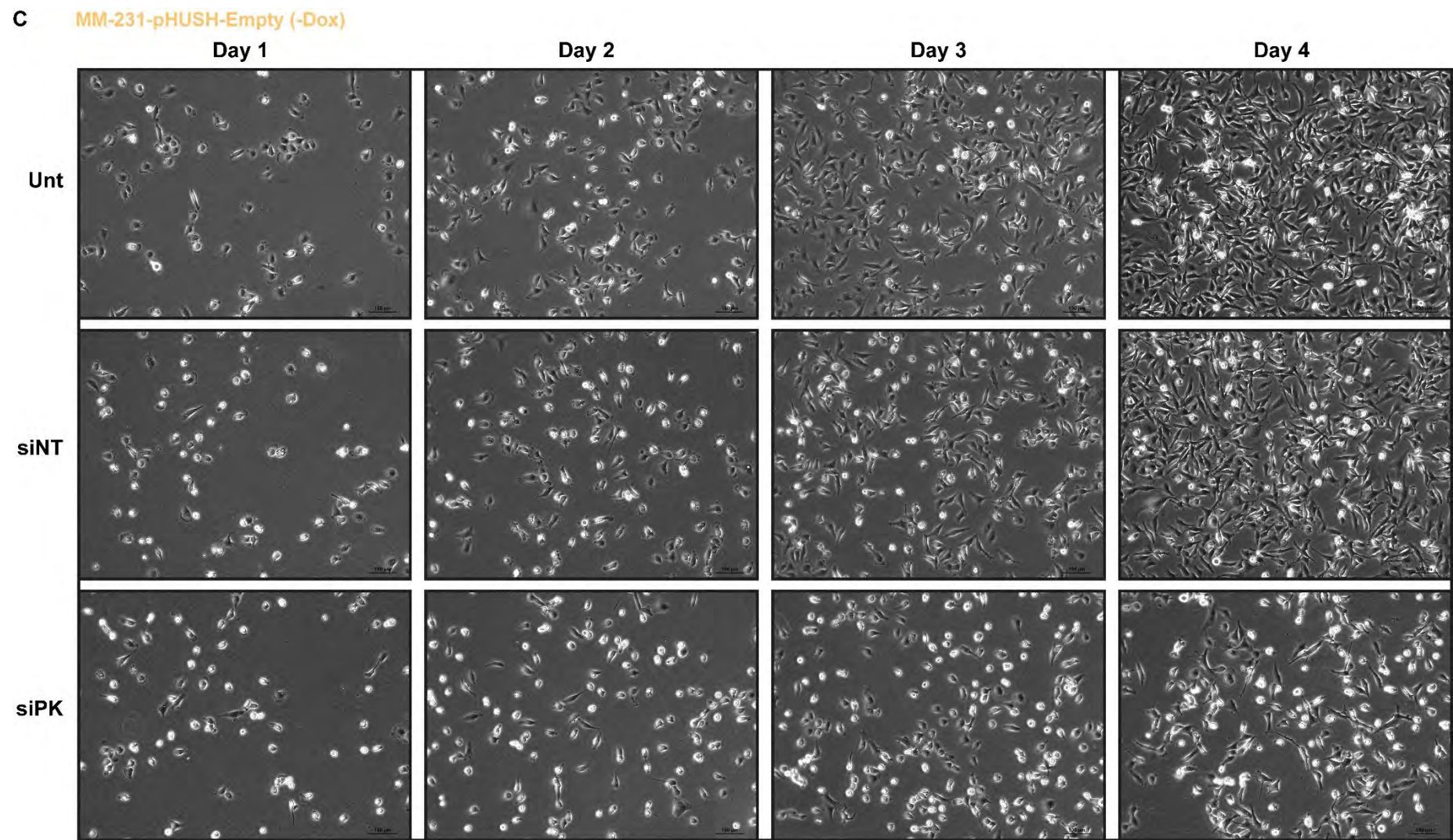




Figure 5.15C: DNA-PKcs knockdown alters cell phenotype over a 4-day timecourse (MDA-MB-231 cells)



**Figure 5.15: DNA-PKcs knockdown alters cell phenotype over a 4-day timecourse**

Representative light microscope images of vehicle-treated MCF7- (A), T47D- (B) and MDA-MB-231- (C) pHUSH-Empty cell lines taken at days 1-4 after siRNA transfection. Images demonstrate the evolving phenotype of DNA-PKcs knockdown in the absence of ELF5 overexpression. Unt, untransfected; siNT, transfection with non-targeting siRNA; siPK, transfection with siRNA against DNA-PKcs.

Images comparing the non-targeting control to DNA-PKcs knockdown at 96 hours (-/+ doxycycline) are shown for each cell line in Figures 5.16A-C. Similar to the results of the cell number analysis, the doxycycline- and vehicle-treated cells had a similar phenotype, indicating that ELF5 overexpression was not having a large impact on the dominant DNA-PKcs knockdown effects. A striking phenotype in the MCF7 and T47D lines with DNA-PKcs knockdown was cells “piling up” on each other, in contrast to the normal growth of these cells in a monolayer. This becomes obvious by 72 hours post-transfection, affecting a small proportion of MCF7 cells and almost all T47D cells. This effect can be more clearly seen in the close-up images shown in Figure 5.17A (MCF7 lines) and Figure 5.17B (T47D lines), with the individual cells within the clump becoming indistinguishable. This phenotype did not occur in the MDA-MB-231 lines.

Additional features of the MCF7 lines with DNA-PKcs knockdown included an increase in the number of “spiky” cells, as well an increase in the number of enlarged, flattened cells or “fried egg” cells (Figure 5.17A, panel v). In addition, there were areas of enlarged and elongated cells growing in a flattened, plate-like manner and featuring very prominent nuclei (Figure 5.17A, panel vi). As the main objective of these experiments was to determine the effect on DNA-PKcs on ELF5 transcriptional activity, the mechanisms behind these phenotypic effects were not explored. The unusual appearance of these cells hints at possible alterations in genomic stability and/or mitotic regulation arising from DNA-PKcs depletion, although this has not been further investigated.

**Figure 5.16: ELF5 overexpression does not affect the knockdown phenotype**

Representative light microscope images taken at day 4 after transfection for MCF7- (A), T47D- (B) and MDA-MB-231- (C) ELF5 and pHUSH-Empty cell lines. Cells were treated with doxycycline (+Dox) or vehicle (-Dox) and transfected with non-targeting siRNA (siNT, left) or DNA-PKcs siRNA (siPK, right). (A) MCF7-ELF5-Isoform2-V5 cells (rows 1 and 2) and MCF7-pHUSH-Empty cells (row 3). (B) T47D-ELF5-Isoform2-V5 cells (row 1), T47D-ELF5-Isoform3-V5 cells (row 2), T47D-pHUSH-Empty cells (row 3). (C) MDA-MB-231-ELF5-Isoform2-V5 cells (row 1), MDA-MB-231-ELF5-Isoform3-V5 cells (row 2), MDA-MB-231-pHUSH-Empty cells (row 3).



Figure 5.16A: ELF5 overexpression does not affect the DNA-PKcs knockdown phenotype (MCF7 cells)

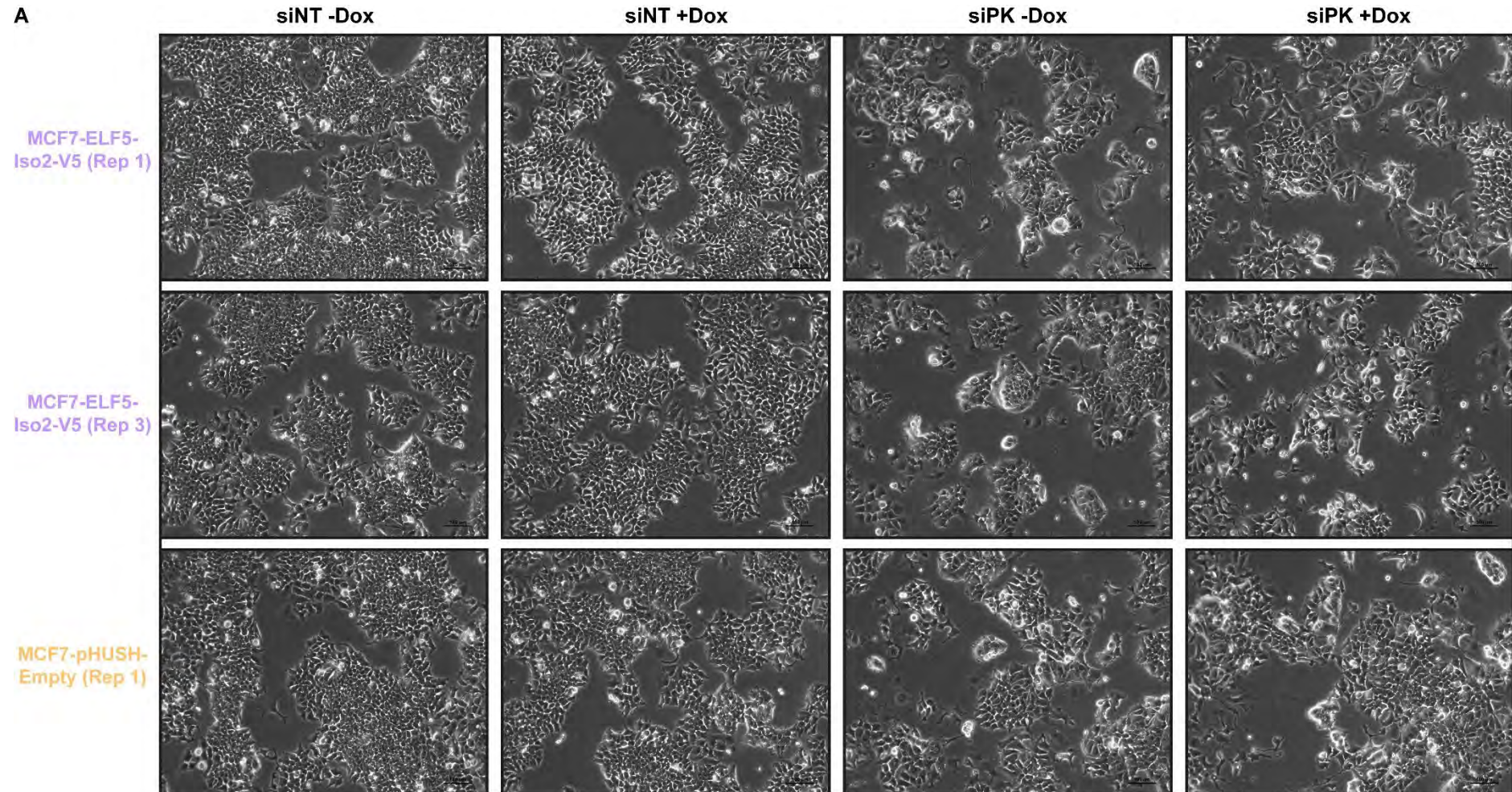
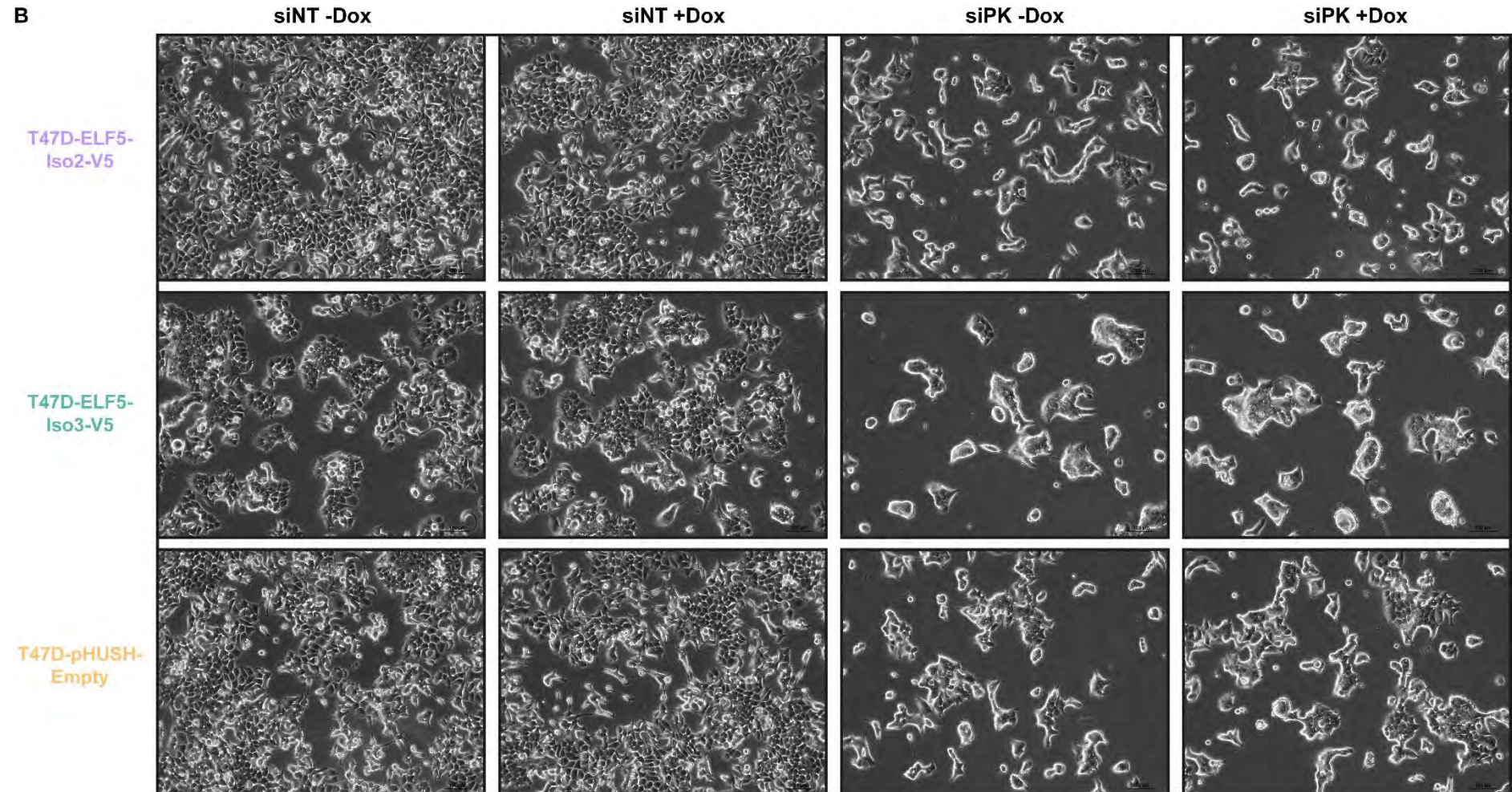


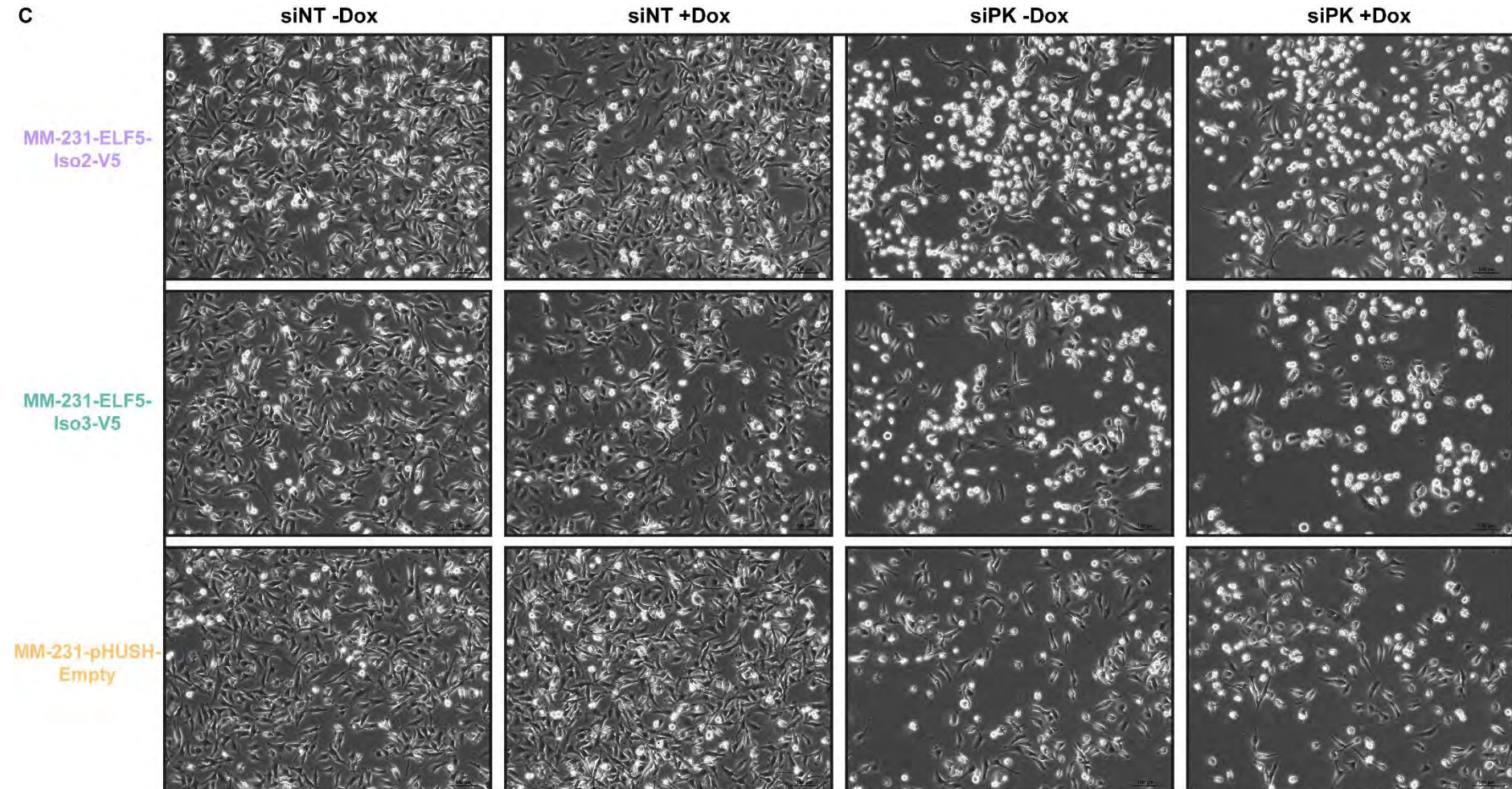


Figure 5.16B: ELF5 overexpression does not affect the DNA-PKcs knockdown phenotype (T47D cells)

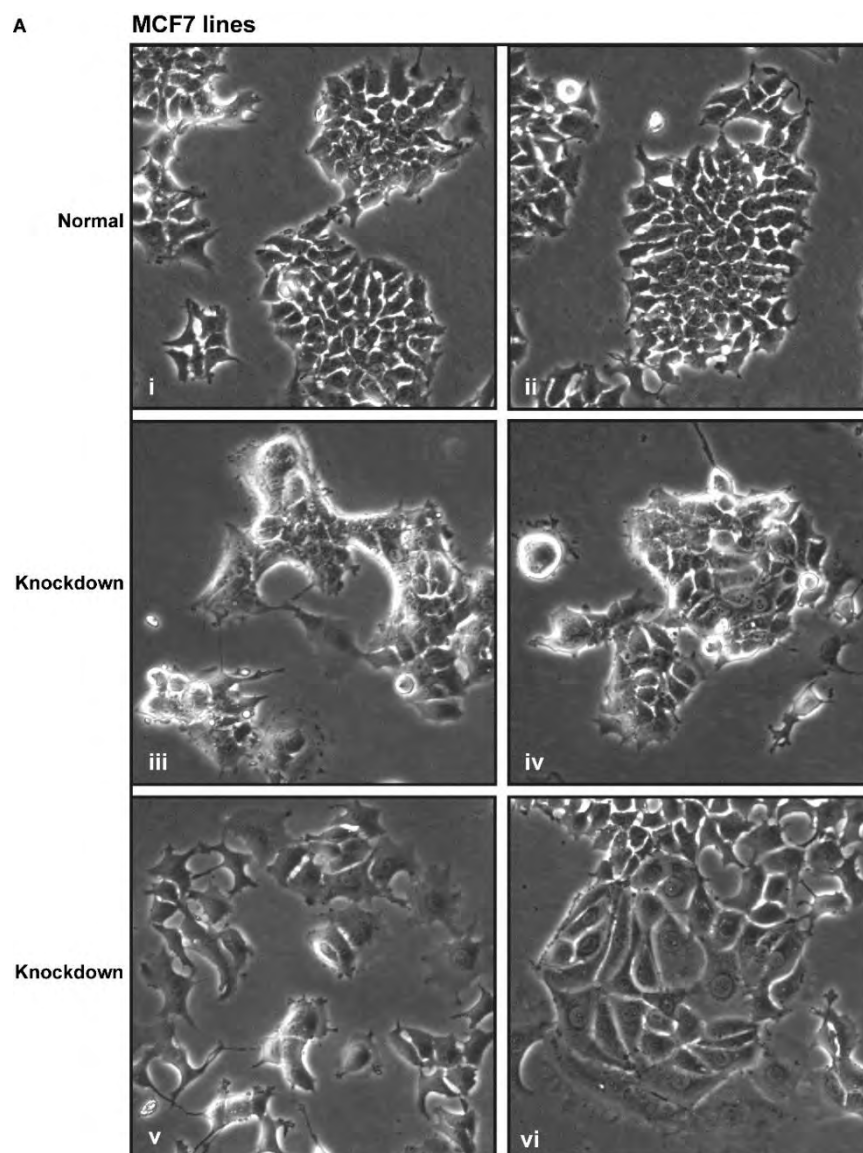




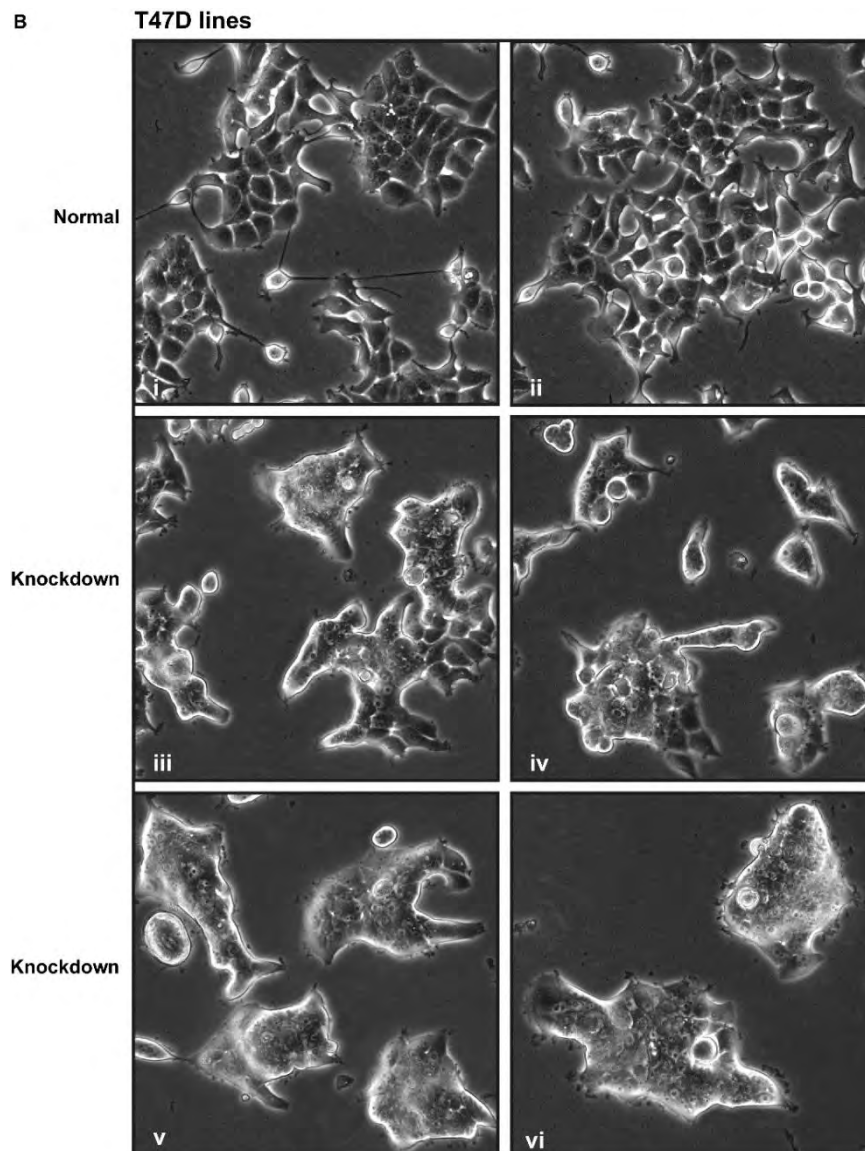
**Figure 5.16A: ELF5 overexpression does not affect the DNA-PKcs knockdown phenotype (MDA-MB-231 cells)**



**Figure 5.17A: Enlarged images of DNA-PKcs knockdown phenotype in luminal breast cancer cell lines (MCF7 cells)**



**Figure 5.17B: Enlarged images of DNA-PKcs knockdown phenotype in luminal breast cancer cell lines (T47D cells)**



**Figure 5.17: Enlarged images of DNA-PKcs knockdown phenotype in luminal breast cancer cell lines**

Close-up light microscope images of MCF7 (A) and T47D (B) luminal breast cancer lines illustrating the various phenotypes observed with DNA-PKcs knockdown at day 4 post-transfection. Images (i) and (ii) in each panel are untransfected cells, while images (iii)-(vi) are DNA-PKcs knockdown cells. (A) Images (i) and (ii) demonstrate the normal MCF7 phenotype, with cells growing in a monolayer, while the knockdown cells in images (iii) and (iv) are piling on top of each other. Image (v) shows increased cellular protrusions (“spikiness”) and enlarged, flattened (“fried egg”) cells. Image (vi) shows an area of large MCF7 cells with prominent nuclei growing in a flattened, plate-like manner. (B) Images (iii)-(vi) demonstrate the main T47D knockdown phenotype, with cells piling on top of each other to the point where individual cells become indistinguishable.

## Effects of DNA-PKcs knockdown on ELF5 transcriptional function (gene expression)

A quantitative PCR (qPCR) panel was compiled to examine the effects of DNA-PKcs knockdown on the ability of ELF5 to regulate known transcriptional targets. The genes for the panel were selected based on a range of criteria, including significant expression changes seen with ELF5 overexpression in previous MCF7-ELF5-V5 microarray and/or RNA-seq experiments (Chapter 4), the presence of an ELF5 ChIP-seq peak in the promoter region of the gene in MCF7-ELF5-V5 cells (Chapter 4), known effects of ELF5 on gene expression in previous qPCR experiments in T47D and MDA-MB-231 lines (Chapter 3) and biological relevance of the target gene. The genes selected for the qPCR panel are shown in Table 2.7 (Chapter 2), which also summarises the selection data for MCF7-ELF5-V5 cells and the quality control data from the qPCR analysis.

The MCF7-, T47D- and MDA-MB-231- ELF5 cell lines were treated according to the previously described knockdown protocol and RNA was collected at 96 hours post-transfection (48 hours of doxycycline treatment). The results of the qPCRs, from three biological replicates, are shown in Figure 5.18. The table above each graph shows the mean calibrated normalised relative quantity values (row 1), fold change (and statistical significance, indicated by colour) between matched doxycycline- and vehicle-treated samples (row 2) and statistical significance for changes occurring as a result of DNA-PKcs knockdown in vehicle-treated cells (row 3) and doxycycline-treated cells (row 4). Some qPCR assays (including *DNA-PKcs*, *ELF5*, *ESR1*, *FOXA1*, *GATA3*, *CDH1* and *VTCN1*) have matching protein expression shown in Figure 5.19A-C.

The induction of ELF5 expression was variable between cell lines (Figure 5.18A). In the T47D lines, the increase in ELF5 expression was relatively small, particularly in the Isoform3 line. This may be related to the length of time these cells had been in culture (passages 7-12), as it has been previously observed that the pooled lines lose ELF5-V5 induction over time. This is due to the baseline leakiness of the vector, resulting in low-level ELF5 expression (particularly in high-expressing inducible cells) and leading to reduced proliferation, increased cell death and cell detachment, and gradual loss of high-expressing cells from the culture. Despite the low level of ELF5 induction, ELF5-V5 was visible by western blot in both ELF5-inducible T47D lines (Figure 5.19B).

In the T47D lines, the knockdown of DNA-PKcs led to a significant increase in the level of ELF5 expression. This affected both inducible ELF5 (Isoform2-V5 and Isoform3-V5 lines) as well as endogenous ELF5 (pHUSH-empty line). A similar trend was also seen

in the MCF7-Isoform2-V5 line, although this did not reach statistical significance. The increase in ELF5-V5 expression with DNA-PKcs knockdown could also be seen at the protein level in the MCF7 and T47D lines (Figure 5.19A-B). No increase in endogenous ELF5 was seen at the protein level in the empty vector lines, although detection is limited by the sensitivity of the ELF5 N-20 antibody. Interestingly, none of the MDA-MB-231 lines showed this increase in ELF5 level and, in fact, at the protein level ELF5 induction was slightly decreased with DNA-PKcs knockdown (Figure 5.19C).

DNA-PKcs knockdown was confirmed by both qPCR (Figure 5.18B) and western blot (Figure 5.19A-C). The knockdown was slightly less efficient (in terms of percentage reduction) in the T47D lines compared to the MCF7 and MDA-MB-231 lines. DNA-PKcs expression was not affected by ELF5 overexpression in any cell line.

There were a variety of effects seen on ELF5 target gene expression in the presence of DNA-PKcs knockdown. The most striking effect was seen in three genes identified as strongly up-regulated in MCF7-ELF5-Isoform2-V5 cells by RNA-sequencing (*PIP*, *VTCN1* and *GRHL3*) (Figure 5.18C-E). The upregulation of all three genes in doxycycline-treated MCF7-ELF5-Isoform2-V5 cells was confirmed by qPCR, with no upregulation of *PIP* or *VTCN1* seen in the empty vector line. There was a trend of increased *GRHL3* expression with doxycycline in the MCF7 empty vector line, which reached statistical significance only in the untransfected cells. Knockdown of DNA-PKcs increased the baseline expression of all three genes in each cell line, with the exception of *PIP* in the empty vector control (which did not express any *PIP*). ELF5 overexpression in combination with DNA-PKcs resulted in additional increases in expression of all three genes. ELF5-induced expression of *PIP*, for example, increased from 14-fold in the untransfected cells to 86-fold in the knockdown cells. Similarly, *VTCN1* induction increased from 12-fold to 24-fold. The baseline expression level of *PIP* and *VTCN1* was relatively low, which may account for the dramatic fold change increases seen with qPCR. However, a similar effect, on a smaller scale, was also seen for the robustly expressed *GRHL3* gene.

In the T47D lines, knockdown of DNA-PKcs caused a similar upregulation of all three genes. Unfortunately, ELF5 induction did not result in any significant expression changes in these genes in the T47D lines, making assessment of the effect of combined knockdown and overexpression impossible. In fact, no genes in the panel were significantly altered by doxycycline treatment in the T47D-ELF5 lines. Similarly, there are no data for these three genes in MDA-MB-231 cells due to very low or absent expression.

A second set of genes (*MATN3* and *SNAI2*) was downregulated by DNA-PKcs knockdown in a cell-type-specific manner (Figure 5.18F-G). Previous studies have demonstrated that these genes are also repressed by ELF5 in breast cancer cell lines (Chakrabarti *et al.*, 2012a; Piggin *et al.*, 2016). This initially suggested a possible knockdown-mediated increase in ELF5 repressive activity, similar to the increase in positive gene regulation seen above. Both *MATN3* and *SNAI2* were indeed downregulated in MCF7-ELF5-Isoform2-V5 cells treated with doxycycline. However, a significant (although smaller) downregulation also occurred in the doxycycline-treated empty vector control cells. Furthermore, no consistent downregulation was seen in the doxycycline-treated T47D- or MDA-MB-231- ELF5 lines, again making assessment of the combined effects of DNA-PKcs knockdown and ELF5 overexpression difficult. DNA-PKcs knockdown (independent of ELF5 overexpression) decreased the baseline expression of these genes in selected cell lines. The expression of *MATN3*, for example, was significantly decreased in the vehicle- and doxycycline-treated MDA-MB-231 knockdown cells (compared to untransfected cells and vehicle-treated non-targeting cells), while the expression of *SNAI2* was significantly decreased in the MCF7 knockdown cells.

A third set of genes (*DKK1*, *FILIP1L* and *LYN*) was oppositely regulated by DNA-PKcs knockdown and ELF5 overexpression (Figure 5.18H-J). All three genes were significantly downregulated by doxycycline treatment in MCF7-ELF5-Isoform2-V5 cells but not in the empty vector control, consistent with the results from previous Affymetrix arrays and RNA-sequencing. These genes were not significantly altered by doxycycline treatment in any of the T47D or MDA-MB-231 lines. DNA-PKcs knockdown, however, resulted in an increase in the expression level of these genes in a cell-type-specific manner. *DKK1* expression, for example, was increased by DNA-PKcs knockdown in a number of different cell lines. Despite the increased baseline expression level of *DKK1* in the MCF7-ELF5-Isoform2-V5 cells, the magnitude of the fold change reduction with ELF5 overexpression was similar to that seen in the absence of DNA-PKcs knockdown. *FILIP1L* expression was significantly increased by DNA-PKcs knockdown in 2 of 3 T47D lines, while *LYN* showed a small but statistically significant increase in all MDA-MB-231 lines. Therefore, DNA-PKcs knockdown and ELF5 overexpression do not affect gene expression in the same direction in all cases.

The final set of genes showed minimal effects of doxycycline treatment and some cell-type-specific (although inconsistent) changes in expression caused by DNA-PKcs knockdown. Genes involved in oestrogen-regulated transcription (*FOXA1*, *GATA3* and



*ESR1*) were unaffected by ELF5 overexpression in the MCF7 and T47D lines (Figure 5.18K-M). In the MDA-MB-231-ELF5 lines, *FOXA1* expression was slightly increased by ELF5 overexpression (although the magnitude was unaffected by DNA-PKcs knockdown), while *GATA3* expression was unchanged. DNA-PKcs knockdown resulted in increased expression of *FOXA1* and *GATA3* expression in the T47D lines only. Interestingly, *ESR1* expression showed a trend towards decreased expression in the MCF7 lines and increased expression in the T47D lines (reaching statistical significance only in the Isoform2-V5 line) with DNA-PKcs knockdown.

Other genes in this final set included *STAT1*, *GDF15* and *CDH1* (Figure 5.18N-P). *STAT1* and *CDH1* expression were upregulated by DNA-PKcs knockdown in the T47D lines only. Similarly, *GDF15* showed a trend towards increased expression in the MDA-MB-231 lines, although the expression of *GDF15* was highly variable. The final gene, *SPDEF*, showed no consistent significant changes with either ELF5 overexpression or DNA-PKcs knockdown (Figure 5.18Q).

The main limitation of this study was the lack of robust ELF5-induced expression changes in gene expression in both the T47D lines and MDA-MB-231 lines. The results from these lines were therefore mainly useful for assessing the cell-line-specific changes in expression caused by DNA-PKcs knockdown, rather than the influence of DNA-PKcs level on ELF5 function. The minimal ELF5-induced expression changes may be related to the low level of ELF5 induction, particularly in the T47D lines. Alternatively, this may be because the ELF5 target genes were primarily selected based on MCF7 data and ELF5 is likely to have cell-type-specific effects on gene expression. To address this, clonal cell lines (with more uniform ELF5 expression) and cell-type-specific ELF5 target genes could be used in future experiments.

The effects of DNA-PKcs level on ELF5 target genes were primarily seen in the MCF7 lines, particularly in the increased up-regulation of *PIP*, *VTCN1* and *GRHL3*. The increased baseline (MCF7 and T47D lines) and ELF5-induced (MCF7 lines only) expression of these genes in DNA-PKcs knockdown cells could be due to several factors. Firstly, DNA-PKcs knockdown caused an increase in *ELF5* expression in both the MCF7 and T47D lines, which could be sufficient to increase the levels of positively-regulated ELF5 target genes. However, the magnitude of the increase in *PIP* and

*VTCN1* expression with ELF5 overexpression in MCF7-ELF5 cells (also seen at the protein level for *VTCN1*) appears out of proportion to this small increase in ELF5 level. Secondly, the knockdown cells are much less confluent than the untransfected and

non-targeting cells, which may affect the expression of some genes indirectly. Thirdly, DNA-PKcs could be a direct inhibitor of ELF5 activity, most likely through ELF5 phosphorylation. Knockdown of DNA-PKcs would therefore relieve this inhibition, increasing the ability of ELF5 to regulate its target genes. Finally, DNA-PKcs may be an indirect inhibitor of ELF5 transcriptional function, for example through the regulation of opposing transcription factors (such as ER) that also regulate these genes. Furthermore, these DNA-PKcs-regulated transcription factors may also alter the expression and activity of ELF5, providing a possible explanation for the increased baseline level of ELF5 in DNA-PKcs knockdown cells. Overall, the results of these experiments indicate that DNA-PKcs level does affect the expression of ELF5-regulated genes and that DNA-PKcs may act as a direct or indirect inhibitor of ELF5 activity. However, these effects are complex and are dependent on both the target gene and cell line.

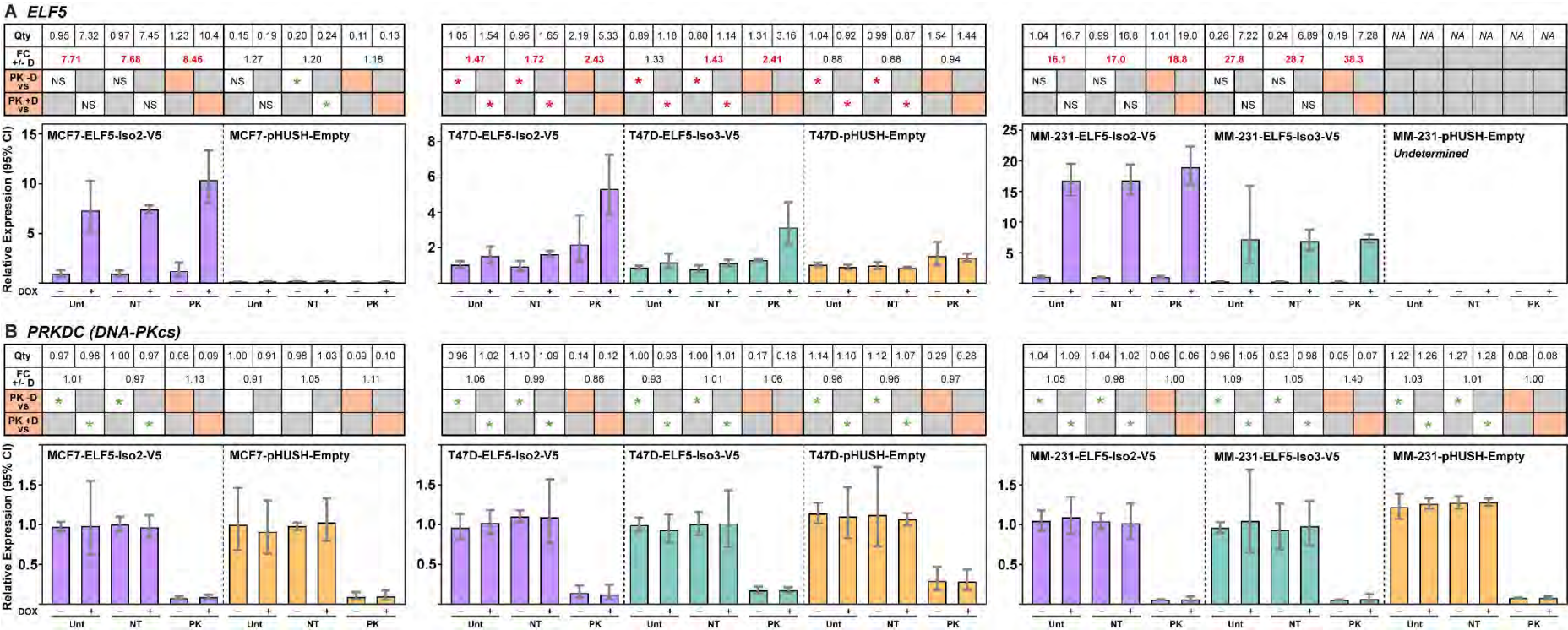
**Figure 5.18: Knockdown of DNA-PKcs affects expression of ELF5-regulated genes**

Quantitative PCR (qPCR) data for selected genes in breast cancer cell lines MCF7, T47D and MDA-MB-231, stably modified with doxycycline-inducible pHUSH-ELF5 isoform 2 or isoform 3 vector (empty vector as a control). Matching protein expression is available for some genes in Figure 5.19. Results for ELF5-Isoform2-V5 and Isoform3-V5 cell lines are shown in purple and green respectively, while empty vector control lines are in yellow. Cells were untransfected (Unt), transfected with a non-targeting siRNA (NT) or transfected with siRNA targeting DNA-PKcs (PK). Cells were also treated with doxycycline (Dox, indicated by + symbol) or vehicle (-). Graphs show the mean calibrated normalised relative quantity values from three biological replicates with 95% confidence interval. For each set of cell lines (shown on the same graph), values are normalised to the ELF5-Isoform2-V5 Unt -Dox sample from the first replicate experiment. The associated table, vertically aligned with the corresponding samples in the graph, provides the exact mean normalised quantity value (Qty, row 1). Row 2 of the table indicates the effects of ELF5 induction on the target gene expression; the fold changes for the vertically aligned +Dox and -Dox sample pairs are shown, with red typeface indicating a significant upregulation (one-way ANOVA), green a significant downregulation and black a non-significant fold change. Rows 3 and 4 of the table indicate the effect of *DNA-PKcs* knockdown on the target gene expression. Row 3 compares the siPK -Dox sample (indicated by the orange box) with each of the Unt -Dox and the siNT -Dox samples, with a red asterisk indicating a significant upregulation and a green asterisk a significant downregulation (NS = no significant difference, one-way ANOVA). Similarly, row 4 compares the siPK +Dox sample (indicated by the orange box) with each of the Unt +Dox and siNT +Dox samples. Assays in the figure are loosely organised by the effects of ELF5 expression and/or DNA-PKcs knockdown. Data are available for only two cell lines in some cases; more detail is provided in Table 5.5. (A) *ELF5* gene expression. (B) *DNA-PKcs* gene expression. (C-E) Genes showing a trend of increased ELF5-induced expression enhanced by DNA-PKcs knockdown (*PIP*, *VTCN1* and



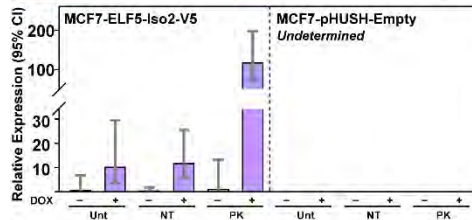
*GRHL3*). (F-G) Genes showing a trend of decreased ELF5- (or doxycycline-) induced expression (*MATN3*, *SNAI2*). (H-J) Genes showing a trend of opposite regulation by ELF5 overexpression and DNA-PKcs knockdown (*DKK1*, *FILIP1L*, *LYN*). (K-M) Genes involved in oestrogen-regulated transcription, minimally affected by ELF5 overexpression or DNA-PKcs knockdown. (N-Q) Genes with no change or variable changes in expression associated with ELF5 overexpression or DNA-PKcs knockdown (*STAT1*, *GDF15*, *CDH1*, *SPDEF*).

Figure 5.18: Knockdown of *DNA-PKcs* affects expression of *ELF5*-regulated genes

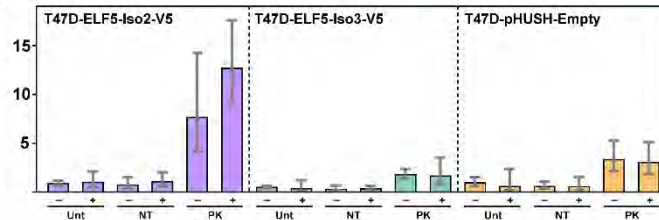


### C PIP

Qty	0.74	10.5	0.56	12.0	1.37	118.1	NA	NA	NA	NA	NA	NA
FC +/- D	14.2		21.4		86.2							
PK-D vs	NS		NS									
PK+D vs	*		*									

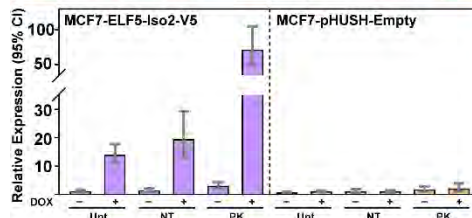


Qty	0.94	1.06	0.79	1.12	7.71	12.7	0.53	0.44	0.35	0.41	1.84	1.70	1.01	0.64	0.64	0.60	3.37	3.10
FC +/- D	1.13		1.42		1.65		0.83		1.17		0.92		0.64		0.94		0.92	
PK-D vs	*		*		*		*		*		*		*		*		*	
PK+D vs	*		*		*		*		*		*		*		*		*	

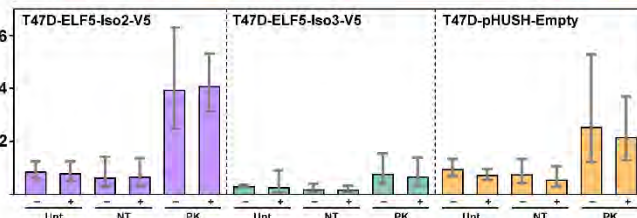


### D VTCN1

Qty	1.15	14.0	1.45	19.6	3.01	71.8	0.73	1.05	1.06	1.04	1.75	2.08
FC +/- D	12.2		13.5		23.9		1.35		0.98		1.19	
PK-D vs	*		*		*		*		*		*	
PK+D vs	*		*		*		*		*		*	

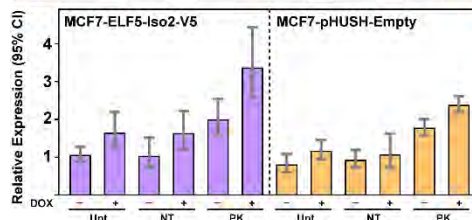


Qty	0.87	0.80	0.64	0.67	3.95	4.09	0.31	0.27	0.20	0.19	0.78	0.68	0.96	0.74	0.76	0.56	2.54	2.17
FC +/- D	0.92		1.05		1.04		0.87		0.95		0.87		0.77		0.74		0.85	
PK-D vs	*		*		*		*		*		*		*		*		*	
PK+D vs	*		*		*		*		*		*		*		*		*	

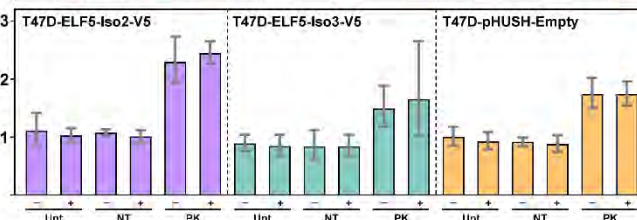


### E GRHL3

Qty	1.06	1.65	1.04	1.63	2.00	3.37	0.81	1.17	0.93	1.08	1.78	2.39
FC +/- D	1.56		1.57		1.69		1.44		1.16		1.34	
PK-D vs	*		*		*		*		*		*	
PK+D vs	*		*		*		*		*		*	

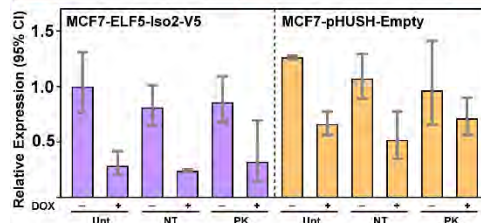


Qty	1.11	1.03	1.08	1.01	2.30	2.45	0.89	0.85	0.84	0.84	1.49	1.65	1.00	0.93	0.92	0.88	1.74	1.74
FC +/- D	0.93		0.94		1.07		0.96		1.00		1.11		0.93		0.96		1.00	
PK-D vs	*		*		*		*		*		*		*		*		*	
PK+D vs	*		*		*		*		*		*		*		*		*	

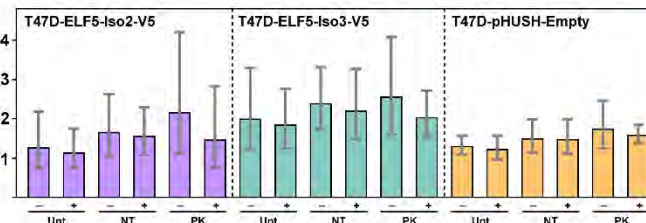


# F MATN3

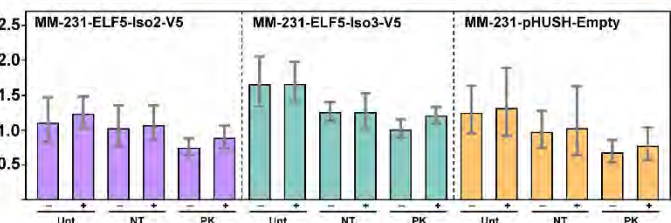
Qty	1.00	0.29	0.81	0.24	0.86	0.32	1.26	0.66	1.07	0.52	0.97	0.71
FC +/- D		0.29		0.30		0.37		0.52		0.49		0.73
PK-D vs	NS		NS				NS		NS			
PK+D vs		NS		NS				NS		NS		



1.28	1.14	1.66	1.57	2.17	1.47	2.00	1.86	2.40	2.21	2.56	2.04	1.31	1.24	1.51	1.49	1.75	1.60
0.89		0.95		0.68		0.93		0.92		0.80		0.95		0.99		0.91	
*		NS				NS		NS				NS		NS			
	NS		NS				NS		NS				NS		NS		

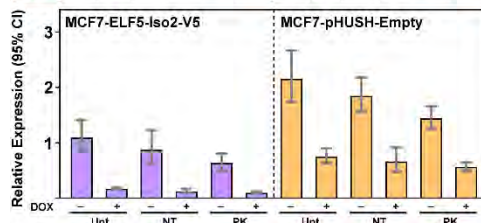


1.11	1.23	1.03	1.07	0.75	0.89	1.66	1.66	1.26	1.26	1.01	1.21	1.25	1.32	0.97	1.02	0.68	0.78
	1.11		0.93		1.19		1.00		1.00		1.20		1.06		1.05		1.15
*		*				*		NS				*		*			
	*		NS			*		NS				*		NS			

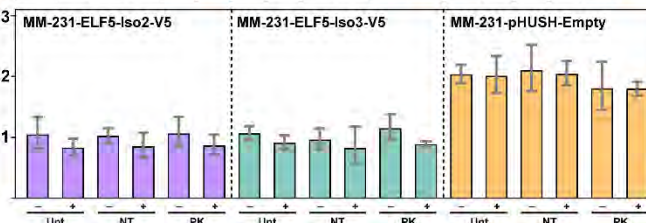


# G SNAI2

Qty	1.10	0.18	0.88	0.12	0.64	0.10	2.15	0.76	1.85	0.66	1.45	0.57
FC +/- D		0.16		0.14		0.10		0.36		0.36		0.39
PK-D vs	*		*				*		NS			
PK+D vs		*		NS			*		NS			



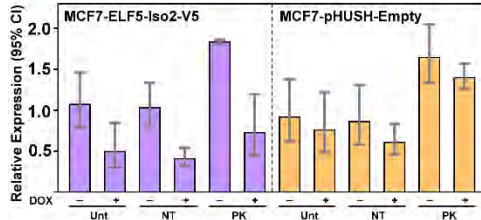
1.05	0.83	1.03	0.85	1.06	0.87	1.07	0.91	0.96	0.83	1.15	0.89	2.04	2.01	2.11	2.05	1.81	1.80
0.79		0.83		0.82		0.85		0.86		0.77		0.99		0.97		0.99	
NS		NS				NS		NS				NS		NS			
	NS		NS				NS		NS				NS		NS		



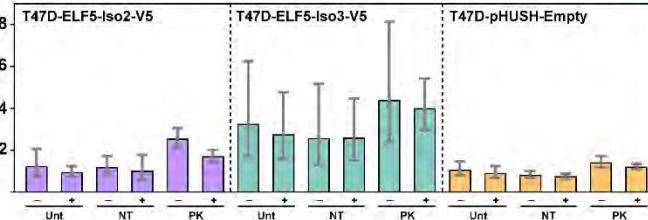


### H DKK1

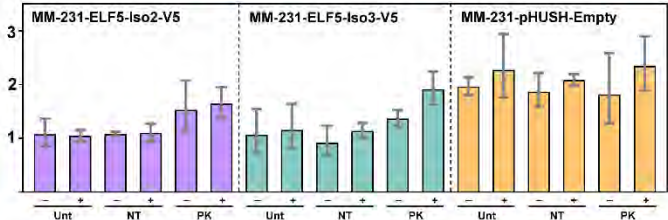
Qty	1.08	0.50	1.04	0.42	1.84	0.73	0.93	0.77	0.87	0.62	1.65	1.41
FC +/- D		0.46		0.40		0.40		0.83		0.71		0.85
PK-D vs	*		*				*		*			
PK+D vs	NS			*			*		*	*		



1.26	0.96	1.20	1.03	2.56	1.73	3.27	2.77	2.60	2.61	4.40	4.02	1.08	0.92	0.83	0.78	1.43	1.23
0.76		0.86		0.68		0.85		1.00		0.91		0.85		0.92		0.86	
*		*				NS		*				NS		*			
*	*	*	*			NS		NS				NS		NS			

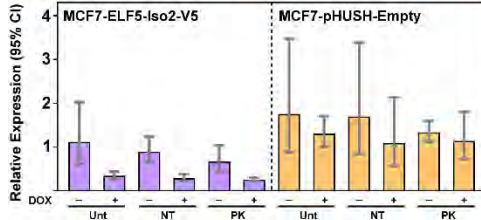


1.08	1.04	1.08	1.10	1.53	1.65	1.07	1.15	0.92	1.14	1.37	1.91	1.96	2.27	1.87	2.08	1.82	2.34
0.96		1.02		1.08		1.07		1.24		1.39		1.16		1.11		1.29	
*		*				NS		*		*		NS		NS			
*	*	*	*				*	*	*	*		NS		NS			

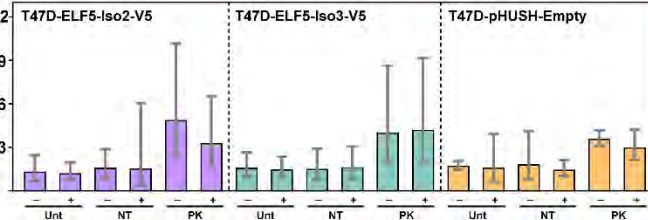


### I FILIP1L

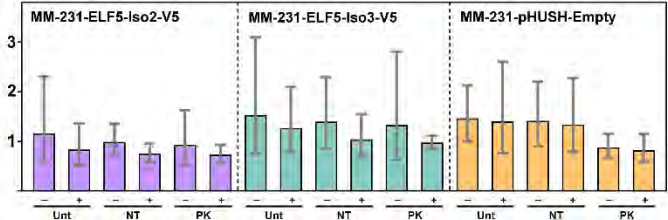
Qty	1.12	0.35	0.90	0.29	0.67	0.26	1.75	1.31	1.69	1.10	1.34	1.14
FC +/- D		0.31		0.32		0.39		0.75		0.65		0.85
PK-D vs	NS		NS				NS		NS			
PK+D vs	NS		NS				NS		NS		NS	



1.32	1.21	1.60	1.55	4.89	3.31	1.60	1.48	1.51	1.62	4.03	4.22	1.74	1.61	1.82	1.48	3.61	3.01
0.92		0.97		0.68		0.93		1.07		1.05		0.93		0.81		0.83	
*		*				*		*				NS		NS			
*	*	*	NS			*	*	*	*			NS		NS			

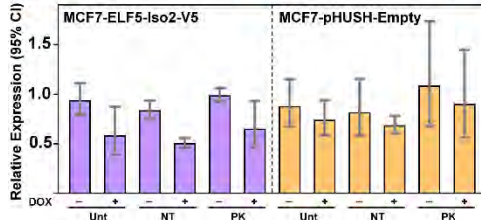


1.16	0.83	0.99	0.75	0.93	0.73	1.53	1.27	1.40	1.04	1.33	0.98	1.46	1.40	1.41	1.34	0.88	0.82
0.72		0.76		0.78		0.83		0.74		0.74		0.96		0.95		0.93	
NS		NS				NS		NS				NS		NS			
NS		NS		NS		NS		NS		NS		NS		NS			

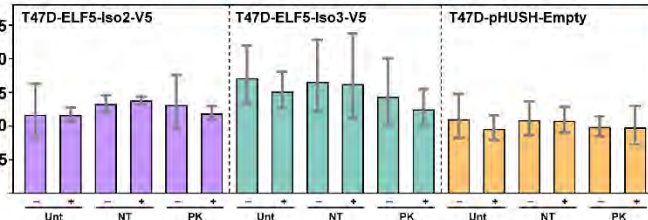


### J LYN

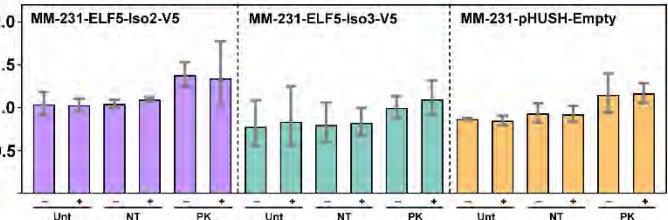
Qty	0.94	0.59	0.84	0.51	0.99	0.65	0.88	0.74	0.82	0.69	1.09	0.90
FC +/- D		0.63		0.61		0.66		0.84		0.84		0.83
PK-D vs	NS		NS				NS		NS			
PK+D vs		NS		NS			NS		NS			



1.17	1.16	1.33	1.38	1.31	1.19	1.71	1.51	1.66	1.63	1.43	1.25	1.10	0.96	1.09	1.08	0.99	0.98
0.99		1.04		0.91		0.88		0.98		0.87		0.87		0.99		0.99	
NS		NS				NS		NS				NS		NS			
NS		NS				NS		NS				NS		NS			

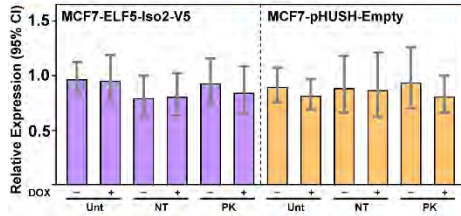


1.04	1.03	1.05	1.10	1.38	1.34	0.78	0.84	0.80	0.82	1.00	1.10	0.87	0.85	0.93	0.92	1.15	1.17
0.99		1.05		0.97		1.08		1.03		1.10		0.98		0.99		1.02	
*		*				*		NS				*		NS			
*	*	NS				*	*	*	*	*		*	*	*	*		

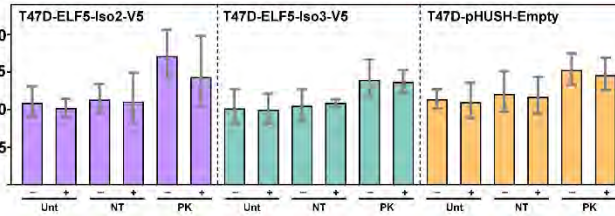


### K FOXA1

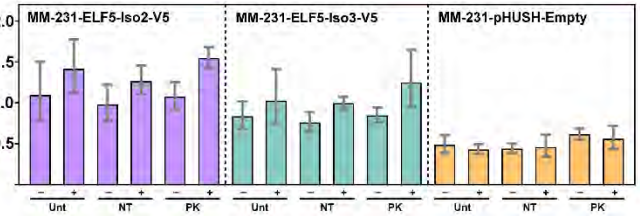
Qty	0.97	0.95	0.79	0.81	0.93	0.94	0.90	0.82	0.89	0.87	0.94	0.81
FC +/- D	0.98		1.03		0.90		0.92		0.98		0.97	
PK-D vs	NS		NS				NS		NS			
PK+D vs	NS		NS				NS		NS			



Qty	1.09	1.02	1.13	1.11	1.71	1.44	1.02	1.00	1.05	1.09	1.40	1.37	1.14	1.10	1.21	1.17	1.53	1.46
FC +/- D	0.94		0.98		0.84		0.98		1.04		0.98		0.96		0.97		0.95	
PK-D vs	*		*				*		*		*		*		NS			
PK+D vs	*		*				*		NS		*		*		NS			

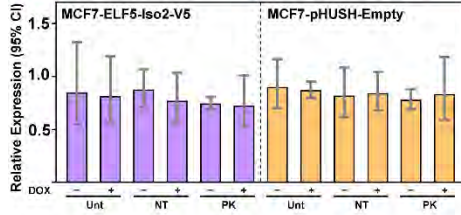


Qty	1.10	1.41	0.98	1.27	1.08	1.55	0.83	1.03	0.76	0.99	0.85	1.25	0.49	0.43	0.44	0.46	0.62	0.56
FC +/- D	1.28		1.30		1.44		1.24		1.30		1.47		0.88		1.04		0.90	
PK-D vs	NS		NS		NS		NS		NS		NS		NS		*			
PK+D vs	NS		NS		NS		NS		NS		NS		NS		*		NS	

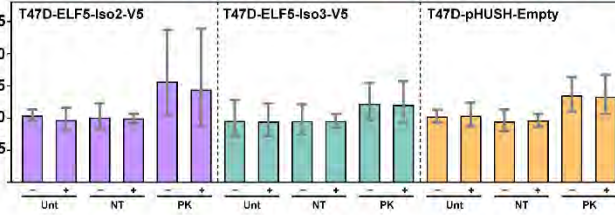


### L GATA3

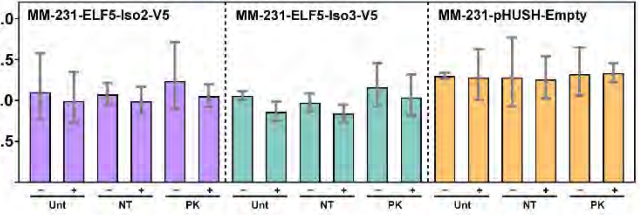
Qty	0.85	0.82	0.88	0.77	0.75	0.73	0.90	0.87	0.82	0.84	0.78	0.83
FC +/- D	0.96		0.88		0.97		0.97		1.02		1.06	
PK-D vs	NS		NS				NS		NS			
PK+D vs	NS		NS				NS		NS			



Qty	1.04	0.97	1.01	0.98	1.57	1.44	0.95	0.95	0.95	0.95	1.22	1.21	1.02	1.04	0.95	0.96	1.35	1.33
FC +/- D	0.93		0.98		0.92		1.00		1.00		0.99		1.02		1.01		0.99	
PK-D vs	*		*				NS		NS				NS		*			
PK+D vs	*		*				NS		NS				NS		*			

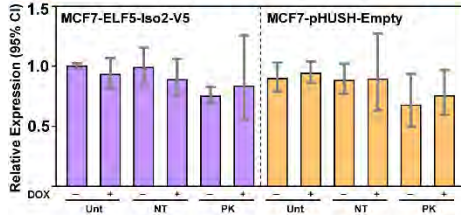


Qty	1.10	1.00	1.08	0.98	1.24	1.05	1.06	0.86	0.97	0.84	1.16	1.04	1.30	1.28	1.28	1.26	1.32	1.34
FC +/- D	0.91		0.92		0.85		0.81		0.87		0.90		0.98		0.98		1.02	
PK-D vs	NS		NS				NS		NS				NS		NS			
PK+D vs	NS		NS				NS		NS				NS		NS			

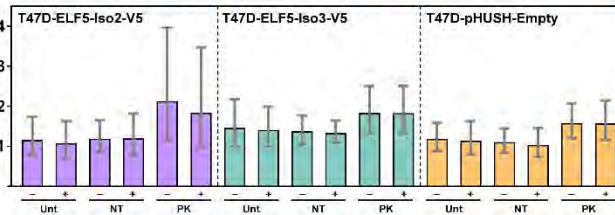


### M ESR1

Qty	1.01	0.94	1.00	0.89	0.76	0.84	0.90	0.95	0.89	0.90	0.68	0.76
FC +/- D	0.93		0.89		1.11		1.06		1.01		1.12	
PK-D vs	*		*				*		*			
PK+D vs	NS		NS				NS		NS			



Qty	1.16	1.08	1.19	1.20	2.13	1.84	1.47	1.41	1.38	1.34	1.83	1.83	1.19	1.14	1.10	1.03	1.58	1.58
FC +/- D	0.93		1.01		0.91		0.96		0.97		1.00		0.96		0.94		1.00	
PK-D vs	*		*				NS		NS				NS		NS			
PK+D vs	*		NS				NS		NS				NS		NS			



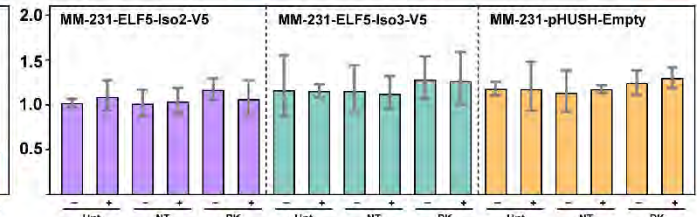
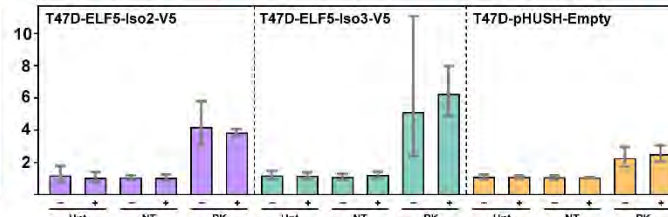
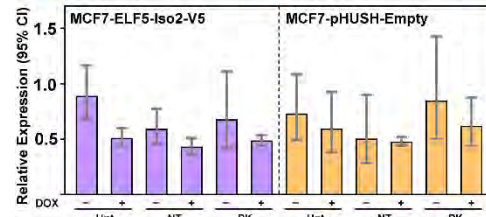


# N STAT1

Qty	0.89	0.51	0.59	0.43	0.68	0.49	0.73	0.60	0.51	0.48	0.85	0.62
FC +/- D	0.57		0.73		0.72		0.82		0.94		0.73	
PK-D vs	NS		NS				NS		*			
PK+D vs		NS		NS			NS		NS			

1.20	1.06	1.07	1.05	4.21	3.87	1.20	1.17	1.10	1.22	5.13	6.25	1.11	1.10	1.07	1.07	2.27	2.51
0.88		0.98		0.92		0.98		1.11		1.22		0.99		1.00		1.11	
*		*				*		*				*		*			
	*		*			*		*				*		*			

1.02	1.09	1.01	1.04	1.17	1.06	1.16	1.16	1.15	1.12	1.28	1.26	1.18	1.18	1.14	1.18	1.24	1.30
1.07		1.03		0.91		1.00		0.97		0.98		1.00		1.04		1.05	
NS		NS				NS		NS				NS		NS			
	NS		NS			NS		NS				NS		NS			

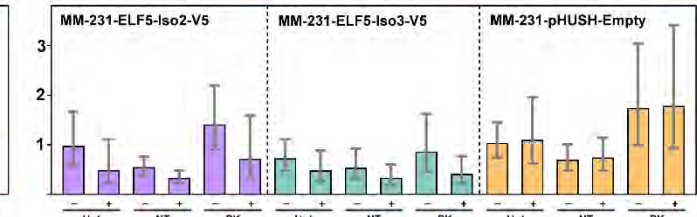
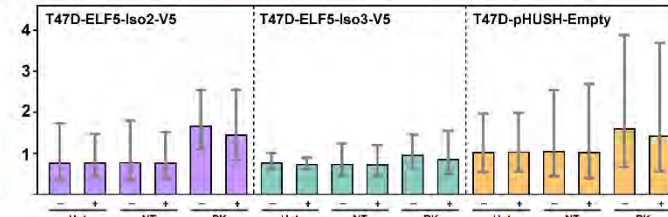
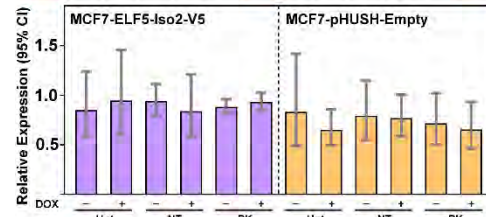


# O GDF15

Qty	0.85	0.95	0.94	0.84	0.88	0.93	0.83	0.65	0.79	0.77	0.72	0.65
FC +/- D	1.12		0.89		1.06		0.78		0.97		0.90	
PK-D vs	NS		NS				NS		NS			
PK+D vs		NS		NS			NS		NS			

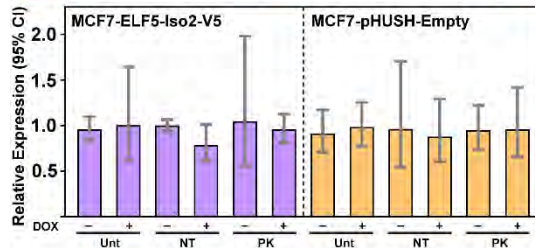
0.78	0.79	0.79	0.78	1.68	1.47	0.79	0.75	0.75	0.74	0.97	0.87	1.04	1.05	1.07	1.04	1.62	1.44
1.01		0.99		0.88		0.95		0.99		0.90		1.01		0.97		0.89	
NS		NS				NS		NS				NS		NS			
	NS		NS			NS		NS				NS		NS			

0.98	0.49	0.55	0.33	1.41	0.72	0.73	0.48	0.54	0.34	0.86	0.41	1.04	1.10	0.70	0.74	1.74	1.78
0.50		0.60		0.51		0.66		0.63		0.48		1.06		1.06		1.02	
NS		*				NS		NS				NS		*			
	NS		*			NS		NS				NS		*			

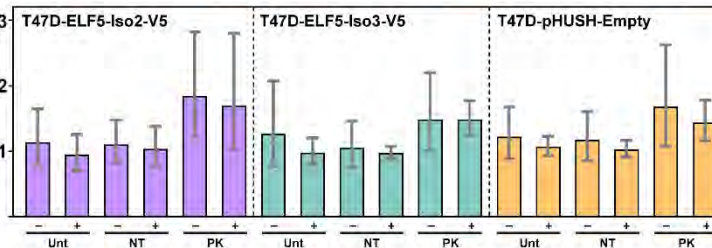


# P CDH1

Qty	0.96	1.01	1.00	0.79	1.05	0.96	0.91	0.99	0.97	0.88	0.95	0.96
FC +/- D	1.05		0.79		0.91		1.09		0.91		1.01	
PK-D vs	NS		NS				NS		NS			
PK +D vs		NS		NS			NS		NS			

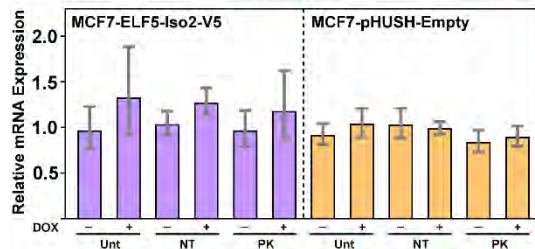


1.14	0.95	1.10	1.04	1.85	1.70	1.27	0.98	1.05	0.98	1.49	1.49	1.22	1.07	1.17	1.03	1.68	1.44
0.83		0.95		0.92		0.77		0.93		1.00		0.88		0.88		0.86	
*		*				NS		NS				NS		NS			
	*		*			*		*				NS		NS			

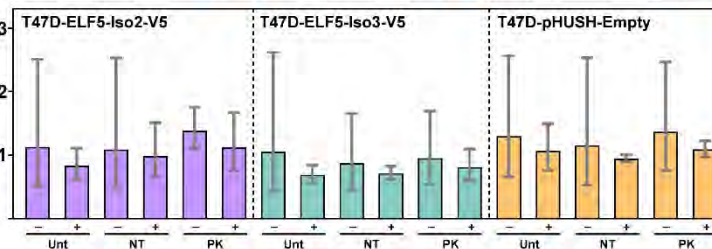


# Q SPDEF

Qty	0.97	1.33	1.04	1.27	0.97	1.18	0.92	1.04	1.03	0.99	0.84	0.90
FC +/- D		1.37		1.22		1.22		1.13		0.96		1.07
PK-D vs	NS		NS				NS		NS			
PK +D vs		NS		NS			NS		NS			



1.13	0.83	1.09	0.99	1.39	1.12	1.05	0.69	0.87	0.72	0.95	0.81	1.30	1.06	1.16	0.95	1.37	1.09
0.73		0.91		0.81		0.66		0.83		0.85		0.82		0.82		0.80	
NS		NS				NS		NS				NS		NS			
	NS		NS			NS		NS				NS		NS			





### **Effects of DNA-PKcs knockdown on ELF5 transcriptional function (protein expression)**

The effects of DNA-PKcs knockdown and ELF5 overexpression were also examined at the protein level (Figure 5.19A-C). DNA-PKcs knockdown was confirmed by western blot in all cell lines. In addition, induction of ELF5-V5 expression was confirmed in all doxycycline-treated ELF5-inducible lines was, along with a small increase in ELF5-V5 expression in doxycycline-treated MCF7- and T47D knockdown cells.

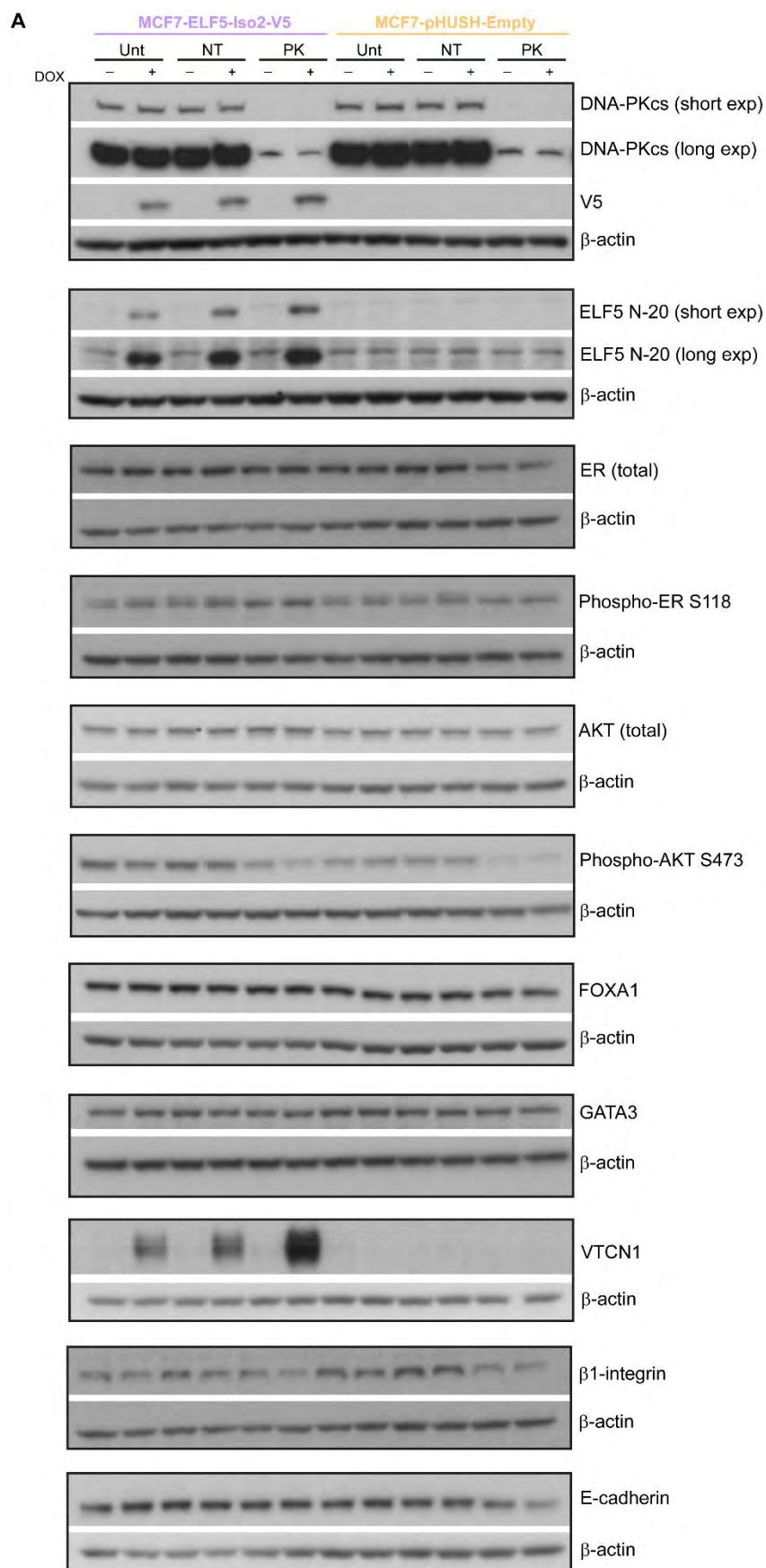
As for the qPCR results, the main effects seen in these blots were caused by DNA-PKcs knockdown, rather than by ELF5 overexpression or the combination of the two. The main exception to this was the large increase in VTCN1 expression that occurred in doxycycline-treated MCF7-ELF5-Isoform2-V5 cells (Figure 5.19A), consistent with the results of the qPCR assay. A small increase in VTCN1 expression was also seen in doxycycline-treated T47D-ELF5-Isoform2-V5 cells, although not in the Isoform3-V5 or empty vector lines (Figure 5.19B).

In all lines, the main effects of DNA-PKcs knockdown were a decrease in AKT phosphorylated at serine 473 (with no change in total AKT) and a decrease in  $\beta$ 1-integrin. While AKT is a known DNA-PKcs substrate,  $\beta$ 1-integrin has not been previously described to be regulated by DNA-PKcs. There were no consistent changes in total ER, phosphorylated ER (serine 118, a known DNA-PKcs phosphorylation site), FOXA1, GATA3 or E-cadherin.

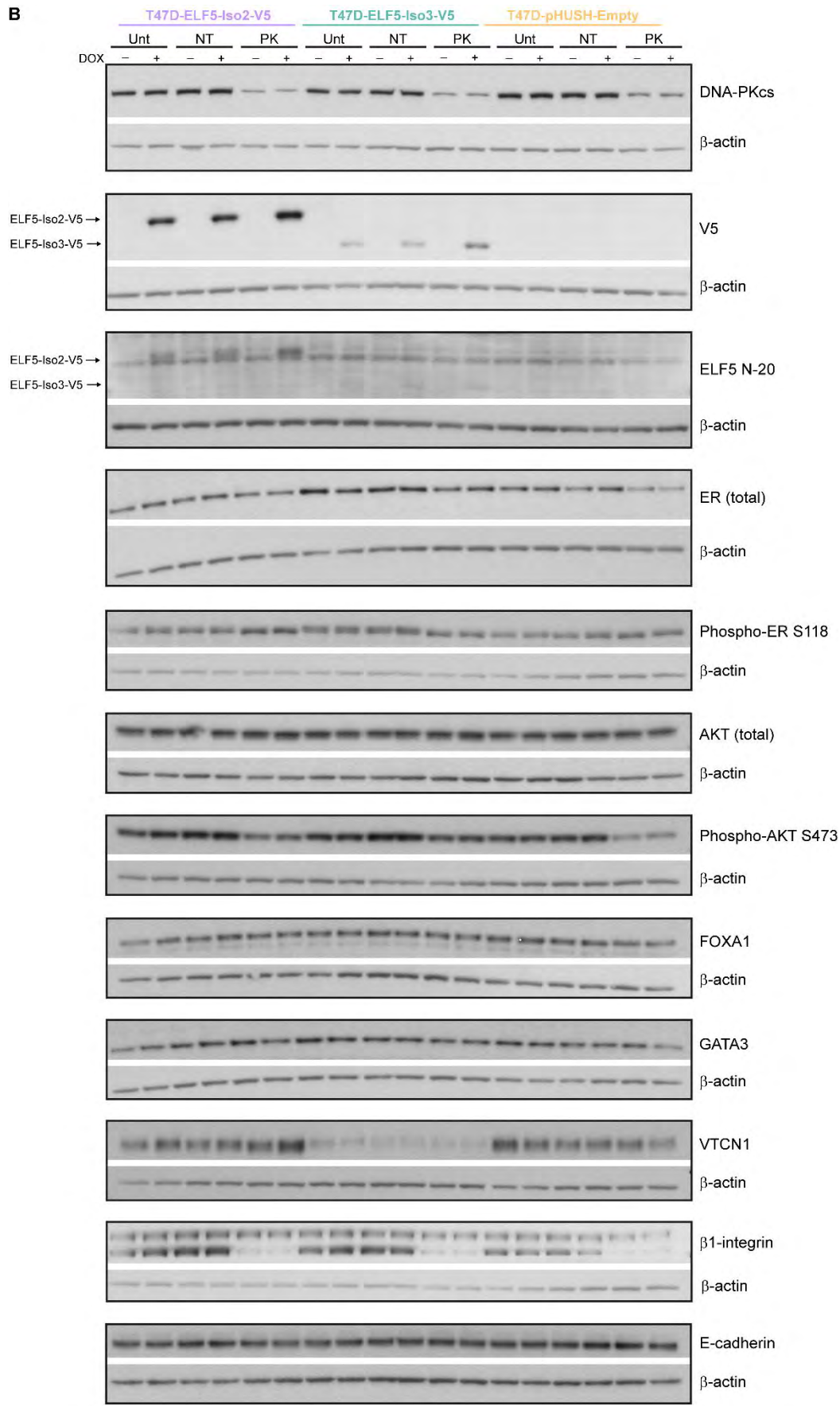
#### **Figure 5.19: DNA-PKcs knockdown affects expression of ELF5 and other breast cancer-associated proteins**

Western blots for MCF7 (A), T47D (B) and MDA-MB-231 (C) cell lines, stably modified with doxycycline-inducible pHUSH-ELF5 isoform 2 or isoform 3 vector (empty vector as a control). Matching gene expression data is available for some proteins in Figure 5.18. Cells were untransfected (Unt), transfected with a non-targeting siRNA (NT) or transfected with siRNA targeting DNA-PKcs (PK). Cells were also treated with doxycycline (Dox, indicated by + symbol) or vehicle (-). Each box represents an individual blot and is shown with the corresponding  $\beta$ -actin loading control.

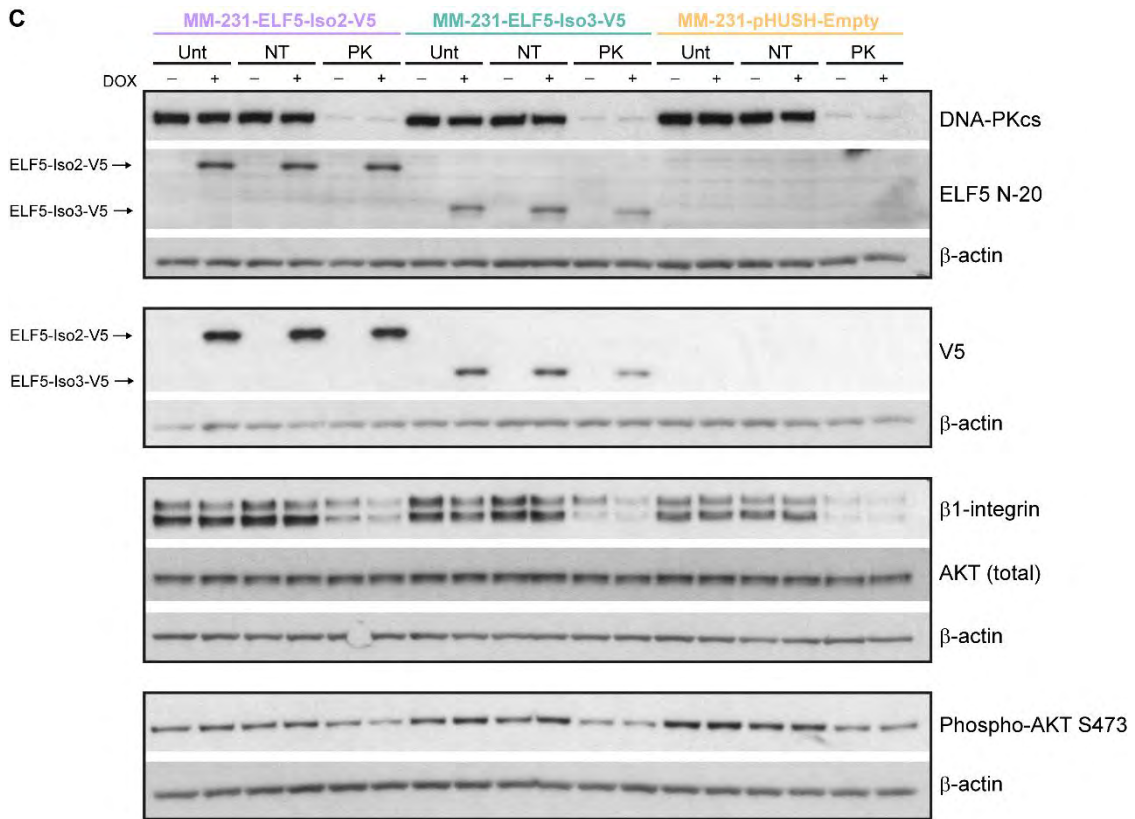
**Figure 5.19A: DNA-PKcs knockdown affects expression of ELF5 and other breast cancer-associated proteins (MCF7 cells)**



**Figure 5.19B: DNA-PKcs knockdown affects expression of ELF5 and other breast cancer-associated proteins (T47D cells)**



**Figure 5.19C: DNA-PKcs knockdown affects expression of ELF5 and other breast cancer-associated proteins (MDA-MB-231 cells)**

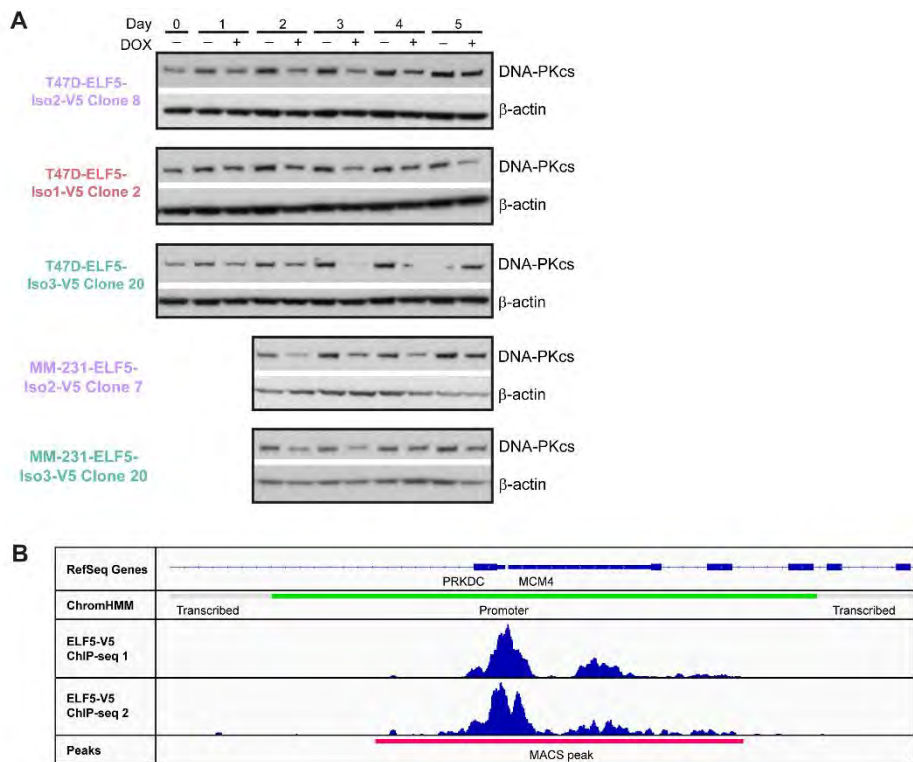


### ELF5 regulation of DNA-PKcs

Many transcriptional regulators function in feedback loops to either limit or reinforce transcriptional effects. DNA-PKcs, for example, phosphorylates ER at serine 118, increasing ER transcriptional activity and preventing ER degradation by the proteasome (Medunjanin *et al.*, 2010b). In addition, DNA-PKcs is a direct transcriptional target of ER (positively regulating DNA-PKcs expression) and is phosphorylated at serine 2056 in response to oestrogen treatment (increasing DNA-PKcs activity) (Medunjanin *et al.*, 2010a). These reciprocal effects establish a positive feedback loop between DNA-PKcs and ER. It was therefore hypothesised that, in addition to being regulated by DNA-PKcs, ELF5 may also regulate DNA-PKcs expression and/or activity.

The pooled cell lines showed very little effect of ELF5 overexpression on DNA-PKcs expression at both the mRNA and protein level (see untransfected samples +/- doxycycline in Figures 5.18 and 5.19). To study this further, DNA-PKcs protein expression was examined in clonal cell lines (used for the ELF5 isoform studies), which produce more uniform and robust ELF5-induced transcriptional effects. These results

demonstrate that ELF5 overexpression (isoforms 1, 2 or 3) is associated with a relative decrease in DNA-PKcs expression (Figure 5.20A). Once again, this may be an indirect effect caused by, for example, the reduction in cell number caused by ELF5 overexpression. However, the identification of a consistent ELF5-V5 ChIP-seq peak in the DNA-PKcs (*PRKDC* gene) promoter in MCF7-ELF5-V5 cells suggests that DNA-PKcs is likely to be a direct ELF5 transcriptional target (Figure 5.20B).



**Figure 5.20: ELF5 decreases DNA-PKcs expression in clonal cell lines**

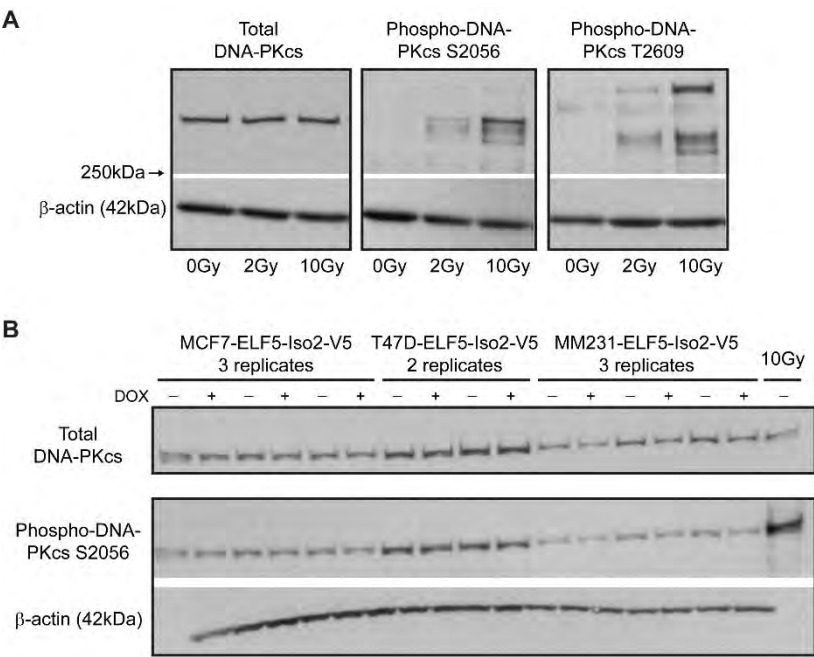
(A) Western blots for DNA-PKcs expression in T47D and MDA-MB-231 clonal cell lines, stably modified with doxycycline-inducible pHUSH-ELF5 vectors (isoforms 1, 2 or 3) and selected for robust ELF5-V5 induction. Cells were treated with doxycycline (Dox, indicated by + symbol) or vehicle (-) and collected every 24 hours from day of plating (T47D lines) or from 48 hours after plating (MDA-MB-231 lines). ELF5-V5 expression in these samples was confirmed by western blot and/or qPCR (see Chapter 3). Due to an experimental issue, the right side of the T47D-ELF5-Isoform3-V5 clone 20 membrane is not clearly visible.

(B) Screenshot from Integrative Genomics Viewer (IGV) showing a reproducible ELF5-V5 ChIP-seq peak in the DNA-PKcs/MCM4 promoter.

The next step was to examine the phosphorylation of DNA-PKcs in response to ELF5 overexpression. Antibodies recognising two distinct DNA-PKcs phosphorylation sites (serine 2056 and threonine 2609) were tested using MCF7-pHUSH-Empty cells treated with ionising radiation (IR) (Figure 5.21A). Only one of these antibodies (S2056)

showed a consistent increase in a band of the correct size with IR treatment. The S2056 antibody was therefore used as a marker of DNA-PKcs auto-phosphorylation and activation in all ongoing studies. Phosphorylation of this site is considered to be a reliable indicator of DNA-PKcs activation (Jette and Lees-Miller, 2015).

As discussed above, the samples from the previous pooled cell line experiments showed little effect of ELF5 overexpression on total DNA-PKcs level. These same samples were re-run on a tris-acetate gel to achieve better resolution of this large protein. Consistent with the previous result, there was no effect of ELF5 overexpression on total DNA-PKcs or phospho-DNA-PKcs, while IR treatment produced a strong increase in phospho-DNA-PKcs (Figure 5.21B). These results suggest that ELF5 does not affect DNA-PKcs phosphorylation. However, given the relatively weak effects of ELF5 induction in the pooled cell lines, repeating this experiment in clonal lines with robust ELF5 expression may provide further insights.



**Figure 5.21: ELF5 does not alter DNA-PKcs phosphorylation in pooled cell lines**

(A) Western blots testing antibodies against DNA-PKcs phosphorylated at serine 2056 (middle) and threonine 2609 (right), with total DNA-PKcs as a control (left). MCF7-pHUSH-Empty cells were untreated (0 Gy) or treated with ionising radiation at a dose of 2 Gy or 10 Gy. (B) Western blots of MCF7-ELF5-Isoform2-V5 (left), T47D-ELF5-Isoform2-V5 (middle) and MDA-MB-231-ELF5-Isoform2-V5 (right) cells treated with doxycycline (Dox, indicated by + symbol) or vehicle (-). An MCF7-pHUSH-Empty sample treated with ionising radiation (10 Gy) is loaded on the far right as a positive control. Samples were run on a tris-acetate gel and blots for total DNA-PKcs and DNA-PKcs phosphorylated at serine 2056 are shown.

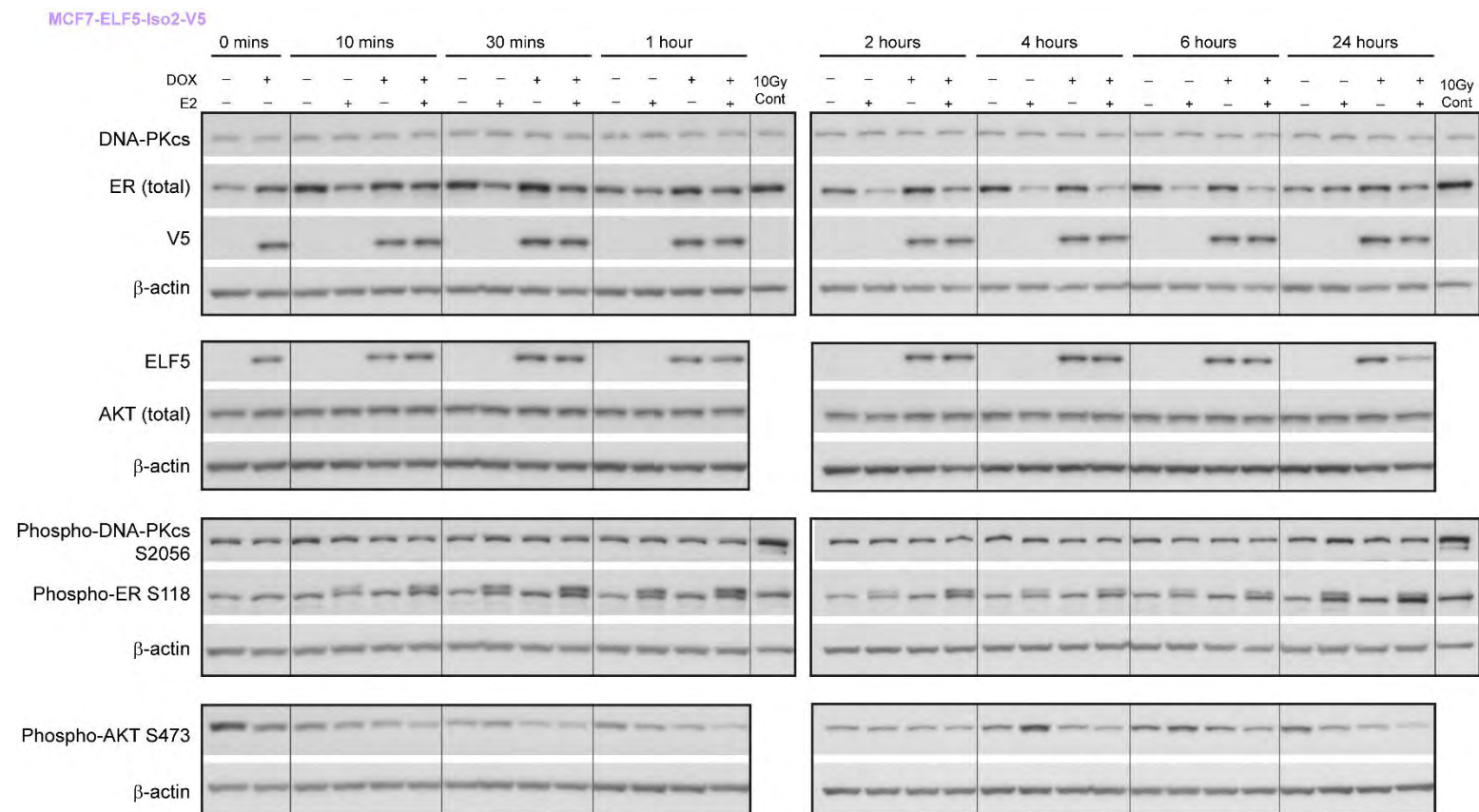
### **Interplay between ELF5, DNA-PKcs and ER**

The results presented so far in this chapter suggest a complex interplay between ELF5, DNA-PKcs and ER. In order to examine this relationship in further detail, MCF7-ELF5-V5 cells were grown in hormone-deprived (HD) medium, consisting of phenol red-free medium supplemented with 10% charcoal-stripped serum. After 24 hours, doxycycline or medium-only treatment was commenced to induce ELF5 overexpression. After 72 hours in HD medium (-/+ 48 hours of doxycycline), the cells were stimulated with oestradiol (E2) or vehicle (ethanol). Cell lysates were then collected at regular intervals for 24 hours to examine the effect of ELF5 overexpression on ER-mediated signalling (Figure 5.22).

E2 treatment resulted in an acute downregulation of total ER, consistent with previous studies demonstrating oestrogen-induced proteasomal degradation of ER (Nawaz *et al.*, 1999). The level of ER downregulation was unaffected by ELF5 overexpression. E2 treatment also induced ER serine 118 phosphorylation, associated with increased ER activity (reviewed in Anbalagan and Rowan, 2015), which once again was not altered by ELF5 overexpression. E2 treatment did not alter ELF5-V5 protein levels at early timepoints, although a decrease in expression was observed at 24 hours in the presence of E2. In addition, total DNA-PKcs and phospho-DNA-PKcs were largely unaffected by either E2 treatment (in contrast to previous studies) or ELF5 overexpression.

An interesting finding from this experiment was the alteration in E2-induced phospho-AKT levels in the presence of ELF5 overexpression. E2 treatment resulted in an increase in AKT phosphorylation at serine 473, seen primarily at the 4- and 6-hour timepoints, with no change in total AKT. Activation of AKT through phosphorylation has been previously described as a rapid, non-genomic ER signalling event (reviewed in Moriarty *et al.*, 2006). Interestingly, phosphorylation at serine 473 has been shown to be catalysed by the mammalian target of rapamycin complex 2 (mTORC2) and DNA-PKcs (reviewed in Bozulic and Hemmings, 2009; Hemmings and Restuccia, 2012) and was shown to be reduced by DNA-PKcs knockdown (Figure 5.19A-C). ELF5 overexpression appeared to prevent this E2-induced increase in serine 473 phosphorylation at the 4- and 6-hour timepoints.





**Figure 5.22: E2-induced ER phosphorylation is not affected by ELF5 overexpression**

Western blots for MCF7-ELF5-Isoform2-V5 cells treated with doxycycline (Dox, indicated by + symbol) or vehicle (-) for 48 hours and stimulated with oestradiol (E2, +) or vehicle (-) for the time indicated. An MCF7-pHUSH-Empty sample treated with ionising radiation (10 Gy) is loaded on the far right as a positive control for DNA-PKcs phosphorylation. Each box represents an individual blot and is shown with the corresponding  $\beta$ -actin loading control.



## DNA-PKcs expression and breast cancer survival

Previous studies examining the relationship between DNA-PKcs expression and survival have produced conflicting results (see Table 1.9, Chapter 1, and associated discussion). However, no studies have examined the association between DNA-PKcs expression and survival in individual molecular subtypes. The impact of DNA-PKcs on survival may involve interactions with uniquely expressed transcription factors, such as ER and ELF5, driving distinct transcriptional programs in different subtypes. Therefore, several platforms, including Km-plotter and cBioPortal, were used to examine the association between DNA-PKcs mRNA expression and survival in each of the molecular subtypes of breast cancer. Results are presented as Kaplan-Meier plots.

Km-plotter (Györfy *et al.*, 2010) utilises microarray expression data from a range of sources including TCGA. Patients were divided by high and low expression of DNA-PKcs using an automatically-generated cut-off within each cohort, with the numbers in each group shown below the graph (Figure 5.23A). Both overall survival (time from diagnosis to death from any cause) and disease-free survival (time to disease relapse) were measured. When all breast cancers were considered together, there was no change in overall survival but a small decrease in disease-free survival (HR=1.38) associated with high DNA-PKcs expression. Similarly, there was a decrease in disease-free survival in the luminal A and luminal B subtypes, with an associated decrease in overall survival in the luminal A subtype only. Interestingly, high DNA-PKcs expression was associated with an increase in overall survival in the HER2-enriched (HR=0.50) and basal-like (HR=0.51) subtypes. However, this was not reflected in the disease-free survival, with the basal-like subtype in fact showing decreased disease-free survival with high DNA-PKcs expression (HR=1.47).

Overall survival was also analysed in the METABRIC cohort, using cBioPortal (Cerami *et al.*, 2012; Gao *et al.*, 2013) to stratify patients by DNA-PKcs expression (Figure 5.23B). Patients were divided by alterations in DNA-PKcs mRNA expression (defined as z-score greater than 2.0 or less than -2.0 compared to the expression distribution for samples diploid for DNA-PKcs). The vast majority of alterations in expression involved DNA-PKcs mRNA upregulation. There were a small number of cases with mRNA downregulation, which were excluded from the survival analysis. Therefore, the plots shown in Figure 5.23B compare patients with DNA-PKcs upregulation (above z-score threshold of 2.0) with patients showing no alteration in DNA-PKcs expression (that is, within the z-score threshold of  $\pm 2.0$ ). In this case, no significant differences in overall survival were identified in either the combined analysis of all breast cancer patients or

in the analysis of individual subtypes. A similar analysis using RNA-sequencing data from the TCGA cohort showed no difference in overall survival or disease-free survival for all breast cancers combined (Figure 5.23C). Survival analysis was not included for the individual subtypes in the TCGA cohort due to the small patient numbers.

Finally, survival was analysed in immunohistochemistry-defined ER-positive and ER-negative patients using Km-plotter (Figure 5.23D) or the METABRIC cohort on cBioPortal (Figure 5.23E). The METABRIC cohort analysis also provides a direct comparison to the DNA-PKcs expression survival analysis performed for ER-positive and ER-negative in a previously published study (Abdel-Fatah *et al.*, 2014). Km-plotter demonstrated a decrease in overall and relapse-free survival associated with high DNA-PKcs expression in ER-positive patients. In contrast, there was a significant increase in overall survival associated with high DNA-PKcs expression in ER-negative patients, with no change in relapse-free survival. In the METABRIC cohort, there was a trend towards decreased survival with upregulated DNA-PKcs expression in the ER-positive group ( $p=0.0530$ ), which disappeared in the ER-negative group (no significant difference in survival,  $p=0.873$ ). A previous study using the METABRIC cohort found a significant increase in survival with high DNA-PKcs expression in ER-negative patients, which was not reproducible here but is consistent with the Km-plotter analysis. However, this study also found a trend towards increased survival with high DNA-PKcs expression in ER-positive patients ( $p=0.061$ ), while the opposite trend was observed in this analysis ( $p=0.053$ ) (Abdel-Fatah *et al.*, 2014). This may be related to use of breast cancer-specific (rather than overall) survival in the 2014 study or the different cut-offs used for defining high and low expression in the two studies. In the 2014 study, for example, 324/437 (74.1%) of ER-negative patients were classified as having high DNA-PKcs expression (software-generated cut-off), while in this analysis 99/433 (22.9%) of ER-negative patients classified as having high DNA-PKcs expression (defined by z-score as above, which is arguably a more objective cut-off). However, the patient numbers in the ER-positive high/low expression groups are similar and it is therefore difficult to account for the discrepancy in results for this group.

The Km-plotter analyses suggest that DNA-PKcs may be influencing survival differently in the various molecular subtypes. The improvement in overall survival seen in the HER2-enriched, basal-like and ER-negative samples with high DNA-PKcs expression contrasts with previous studies suggesting that high DNA-PKcs expression is associated with poorer response to DNA-damaging therapies. However, this survival difference was not seen in the cBioPortal analysis of the METABRIC data. In addition,

the combination of improved overall survival and poorer disease-free survival in the basal-like subtype with high DNA-PKcs expression is difficult to reconcile. One possible explanation may be that high DNA-PKcs expression promotes a longer latency between relapse and death. Overall, these data hint at a possible difference in DNA-PKcs action in breast cancer subtypes, with distinct effects on outcome. However, no definite conclusions could be drawn due to the discrepancies between different analyses as well as several additional factors that will be further addressed in the discussion section of this chapter.

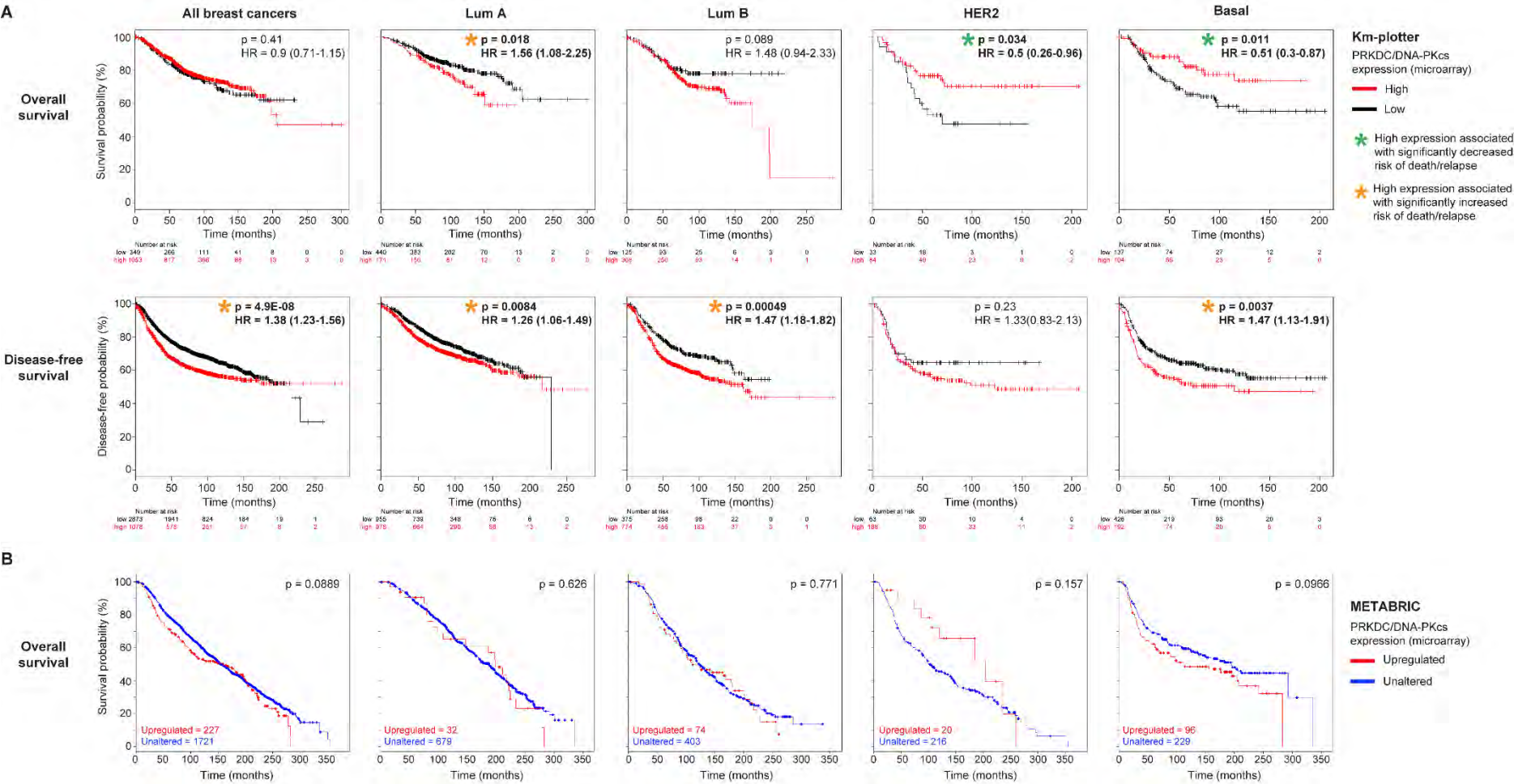
**Figure 5.23: DNA-PKcs expression may have subtype-specific effects on breast cancer survival**

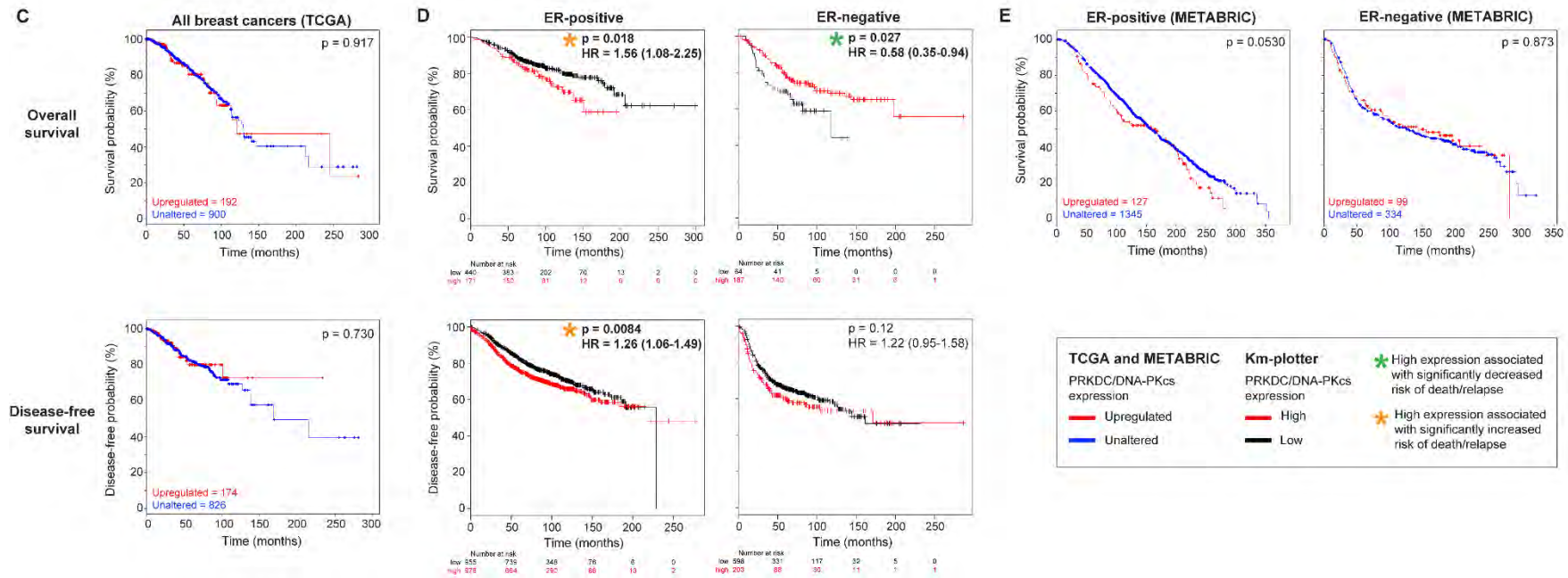
**(next page)**

(A) Kaplan-Meier plots generated using DNA-PKcs microarray expression data from Km-plotter, with overall survival (OS) on the top and disease-free survival on the bottom. The first plot shows all breast cancers combined, followed by individual analysis of luminal A (Lum A), luminal B (Lum B), HER2-overexpressing (HER2) and basal-like (Basal) cancers. Patients are divided into high (red) and low (black) DNA-PKcs expression by an automatically generated cut-off in each sub-group. The total number of patients in the high and low groups are shown in the table (first column) below each plot. P-values were generated by Km-plotter using the log-rank test. Significant p-values (<0.05) are shown in bold font and are marked with an asterisk - green indicates increased survival with high DNA-PKcs expression, while orange indicates decreased survival with high DNA-PKcs expression. The hazard ratio (HR) is also shown for each group. (B) Kaplan-Meier plots generated with cBioPortal using DNA-PKcs microarray data from the METABRIC study showing overall survival for all breast cancer cases and sub-groups as above. Patients are divided into altered (red) and unaltered (blue) DNA-PKcs expression according to z-score, with the total number of patients in each group shown in the bottom left corner. Altered expression is defined as a z-score with an absolute value above 2.0; the majority of cases of altered DNA-PKcs involved DNA-PKcs upregulation (positive z-score) and the small number of cases with DNA-PKcs downregulation (negative z-score) were excluded from the analysis. P-values were generated by cBioPortal using the log-rank test, with significant differences indicated as above. (C) Kaplan-Meier plots generated with cBioPortal using DNA-PKcs RNA-seq data from TCGA. Overall survival (top) and disease-free survival (bottom) are shown for all breast cancer cases. Patients are divided into altered (red) and unaltered (blue) DNA-PKcs expression as for panel B, with the total number of patients in each group shown in the bottom left corner. P-values were generated by cBioPortal using the log-rank test. No subtype-specific analysis was performed for this dataset due to small patient numbers. (D) Kaplan-Meier plots generated using DNA-PKcs microarray expression data from Km-plotter, with overall survival on the top and disease-free survival on the bottom. Breast cancer cases are separated into ER-positive (left) and ER-negative (right) on

the basis of immunohistochemistry. Within each sub-group, patients are divided into high (red) and low (black) DNA-PKcs expression by an automatically generated cut-off. The total number of patients in the high and low groups are shown in the table (first column) below each plot. P-values were generated by Km-plotter using the log-rank test with significant differences indicated as above. (E) Kaplan-Meier plots generated with cBioPortal using DNA-PKcs microarray data from the METABRIC study showing overall survival for ER-positive (left) and ER-negative (right) breast cancer cases. Patients are divided into altered (red) and unaltered (blue) DNA-PKcs expression as in panel B, with the total number of patients in each group shown in the bottom left corner. P-values were generated by cBioPortal using the log-rank test.

Figure 5.23: DNA-PKcs expression may have subtype-specific effects on breast cancer survival





## Discussion

### Overview

Despite the importance of co-operative interactions in transcription factor function, there have been no previous studies investigating ELF5 chromatin binding partners in humans. The results presented in this chapter describe the first use of a mass spectrometry-based method (RIME) to identify novel ELF5-interacting proteins in human breast cancer cells. DNA-dependent protein kinase catalytic sub-unit (DNA-PKcs) was identified as a strong candidate, appearing in all ELF5 RIME experiments and none of the controls. Several known protein partners of DNA-PKcs, functioning in both DNA repair and transcription, were also identified, including Ku80, TOP2A, TOP2B and PARP1. The interaction between ELF5-V5 and DNA-PKcs was confirmed using co-immunoprecipitation of cross-linked samples as well as an immunofluorescence-based proximity ligation assay. ELF5 was also identified as a phosphoprotein *in vivo*. Depletion of DNA-PKcs using siRNA demonstrated that DNA-PKcs regulates ELF5 transcriptional activity in a gene- and cell type- specific manner. DNA-PKcs also regulates ELF5 expression (at both the mRNA and protein level), while ELF5, in turn, regulates DNA-PKcs expression, most likely through direct binding to the DNA-PKcs promoter. Finally, the clinical relevance of DNA-PKcs expression was examined through survival analysis, which suggested subtype-specific roles for DNA-PKcs expression in breast cancer survival; however, due to inconsistencies in the results, no definite conclusions could be drawn. It is hypothesised that the differences in subtype-specific survival relate to the divergent transcriptional functions of DNA-PKcs in the molecular subtypes of breast cancer.

### Advantages and limitations of RIME

Rapid immunoprecipitation of endogenous proteins (RIME) was selected to study the ELF5 interactome in breast cancer cells, as it presents several advantages over other methods (Mohammed *et al.*, 2013; Mohammed *et al.*, 2016). Firstly, the incorporation of a cross-linking step in RIME preserves low-affinity and transient interactions, which are common in dynamic transcriptional complexes. Cross-linking also allows for the use of stringent wash steps, which are important to reduce non-specific binding of proteins to the beads or antibody, but which might otherwise disrupt fragile interactions. Importantly, formaldehyde is a small molecule which only cross-links proteins that are immediately adjacent. RIME may also be used to purify protein complexes without the need to over-express the target protein, although this depends on the availability of a

high-affinity and high-specificity antibody. Unfortunately, an antibody meeting these criteria is not available for ELF5 and the method was therefore modified for inducible ELF5-V5 overexpression in MCF7 stable cell lines.

The RIME method also has some limitations, which need to be considered in the experimental design and analysis. Firstly, the use of a tagged, over-expressed protein, as was required in this case, can lead to the identification of interactions that are not physiologically relevant, due to high protein levels or tag-induced changes in protein structure. In order to minimise these effects, the V5 tag that was selected is small (14 amino acids with 9 amino acid linker sequence) and has been previously demonstrated to have no effect on the ability of ELF5 to regulate its target genes (Kalyuga *et al.*, 2012). A second limitation is that individual peptides with formaldehyde-induced cross-links or lysine modifications will not be identified by the mass spectrometry analysis, which can result in lower sequence coverage compared to non-cross-linked samples. Thirdly, RIME is not able to distinguish discrete transcriptional complexes; it is unknown, for example, whether DNA-PKcs and Ku80, which were identified together in 3 experiments, co-localise with ELF5 in a single complex or exist in individual ELF5 complexes at different genomic locations. Finally, despite the use of cross-linking and stringent washes, RIME and other affinity purification mass spectrometry experiments almost always identify a large number of non-specific proteins. A typical RIME experiment, for example, is expected to identify 300-900 proteins, of which only 5-10% will be specific interactors (Mohammed *et al.*, 2016). These non-specific interactions may be due to binding to the epitope tag, antibody or the affinity matrix (in this case, protein A/G magnetic beads). An isotype-matched antibody control experiment was therefore performed for every ELF5-V5 RIME experiment and a stringent subtractive method was used to exclude any protein from the ELF5-V5 set that appeared in one or more control experiments.

The CRAPome database is a resource that has been developed to address this challenge of differentiating specific and non-specific interactors, compiling negative-control purifications from multiple experiments that use a range of cell lines, epitope tags and affinity matrices (Mellacheruvu *et al.*, 2013). Using this database, DNA-PKcs was identified as a non-specific interactor in a total of 43% (175/411) of all negative-control experiments; this proportion increased to 63% (24/38) when experiments using magnetic beads only were considered. This does indicate a propensity for DNA-PKcs to be purified as a non-specific interacting protein. Given this information, a completely different method, not reliant on affinity purification, was chosen for validation of the



ELF5-V5 and DNA-PKcs interaction. The proximity ligation assay instead utilises the spatial proximity of two interacting proteins to generate a fluorescent signal in fixed cells and was used successfully with two combinations of antibodies to confirm the specific interaction between ELF5-V5 and DNA-PKcs.

### **ELF5 is a phosphoprotein *in vivo***

This is the first study to demonstrate phosphorylation of ELF5 in human cells. Multiple members of the ETS transcription factor family are known to be regulated by phosphorylation (Charlot *et al.*, 2010). Of the closely related epithelial-specific ETS factors (including ELF3, ELF5 and EHF), only ELF3 has previously shown to be phosphorylated; in this case, p21-activated kinase 1 (PAK1) was shown to phosphorylate serine 207 of ELF3 (within a serine and aspartic acid-rich (SAR) domain between the PNT and ETS domains), resulting in increased ELF3 transcriptional activity and protein stability (Manavathi *et al.*, 2007). While ELF5 does not have a SAR domain, many of the predicted ELF5 phosphorylation sites are clustered in an area that includes the end of the Pointed domain and the region between the Pointed and ETS domains (amino acids 93-159 of Isoform 2, Figure 5.11A). The Pointed domain of ELF5 has been shown to have transactivation activity, which suggests phosphorylation of this region may modify the transcriptional activity of ELF5 (Choi and Sinha, 2006). Conversely, the N-terminal region of ELF5 has been shown to have an inhibitory effect on ELF5 transcriptional activity (amino acids 1-42 of Isoform 1 or 1-32 of Isoform 2) (Oettgen *et al.*, 1999). A number of phosphorylation sites are also predicted for this region, suggesting that phosphorylation might either relieve or promote ELF5 auto-inhibition to regulate transcriptional activity. Phosphorylation of specific residues in ETS1, for example, reinforces an auto-inhibitory conformation (Hollenhorst *et al.*, 2011).

The study of ELF5 phosphorylation was stimulated by the discovery of DNA-PKcs as an ELF5-interacting protein. Interestingly, phosphosite prediction revealed that the top ELF5 phosphorylation site, T96 within the Pointed domain, was likely to be catalysed by DNA-PKcs. However, siRNA-mediated depletion of DNA-PKcs did not significantly affect the level of phosphorylated ELF5. This may indicate that DNA-PKcs does not phosphorylate ELF5 or, perhaps more likely, that DNA-PKcs is one of a number of kinases that phosphorylate ELF5. ELF5 Isoform 3 was also shown to be a phosphoprotein, which, while not excluding phosphorylation within the Pointed domain, does indicate the presence of phosphorylation sites outside the Pointed domain. Using less stringent settings, a total of 15 possible ELF5 phosphorylation sites were

predicted, catalysed by up to 23 different kinases. Comparing these predictions to the ELF5 RIME data, the only overlap was casein kinase II subunit alpha (CSK21), identified in a single RIME replicate and predicted to phosphorylate ELF5 at serine 129 between the PNT and ETS domains. Indeed, CK2 is known to phosphorylate numerous transcription factors, including SNAI1, FOXC2 and the ETS factors SPI1 and SPIB, and is important in the maintenance of the epithelial phenotype in breast cancer cells (reviewed in Filhol *et al.*, 2015; Tootle and Rebay, 2005). Interestingly, however, S129, like T96 is absent in ELF5 Isoform 3, indicating the additional involvement of other sites and/or kinases in ELF5 phosphorylation. Many ETS factors are known to be phosphorylated by mitogen-activated protein kinases (MAPKs) and the presence of a MAPK3 site in the ETS domain provides another interesting candidate for future investigations (Tootle and Rebay, 2005).

The role of phosphorylation in the regulation of ELF5 is currently unclear. Phosphorylation, which increases the negative charge of the modified residue, can alter protein conformation, stability, co-factor interactions, subcellular localisation and DNA binding affinity of transcription factors (Filtz *et al.*, 2014). The upstream signalling pathways, converging on activated kinases that catalyse ELF5 phosphorylation, are also currently unknown. ELF5 has previously been shown to be located in both the nucleus and the cytoplasm in ER-positive breast cancers and it is possible that phosphorylation regulates ELF5 subcellular localisation (Gallego-Ortega *et al.*, 2015).

A recent study quantifying the proteome and phosphoproteome for 77 TCGA breast cancer samples was interrogated to examine ELF5 phosphorylation in clinical samples (Mertins *et al.*, 2016). ELF5 was not identified as a phosphoprotein in this study; however, non-phosphorylated ELF5 protein was only detected in 5 of the 77 samples (6.5%), which is most likely related to the low relative abundance. As discussed by the study, “absence calls” are difficult to make using mass spectrometry-based proteomics and the lack of observed peptides does not necessarily mean a protein is not present in the sample. As ELF5 protein expression is already likely to be low in the majority of breast cancers, the mass spectrometry detection of ELF5 phosphopeptides, which has an even lower sensitivity, is unlikely to be successful without the use of a specific purification step.

### **DNA-PKcs as a modulator of ELF5 transcriptional activity**

The identification of a DNA repair protein as a transcriptional co-regulator is not as unusual as it may first seem. As discussed in Chapter 1, the DNA repair and

transcriptional machineries are intimately connected. This relationship may have evolved from the inherently DNA-damaging nature of transcription, which causes torsional stress, exposes single-stranded DNA to potential genotoxic insults and promotes recombination events. The subsequent evolution of this relationship has seen the enzymatic capabilities of many DNA repair proteins (which include glycosylases, helicases, nucleases and kinases) also being utilised in transcriptional regulation (reviewed in Fong *et al.*, 2013).

DNA-PKcs interacts with and regulates a wide variety of transcription factors, including a number of ETS family members (see Table 1.8, Chapter 1) (Brenner *et al.*, 2011; Choul-li *et al.*, 2009). Other proteins also frequently identified in DNA-PKcs transcriptional complexes include those with known functions in both DNA repair and transcription, such as Ku70 and Ku80 (the regulatory sub-units of the DNA-PK complex, which recruit DNA-PKcs to double-stranded DNA breaks), poly(ADP-ribose) polymerase 1 (PARP1) and the DNA topoisomerases 2-alpha and 2-beta (TOP2A, TOP2B). A number of these proteins were also identified in the ELF5 RIME experiments, although less robustly than DNA-PKcs. However, no DNA repair proteins functioning in the downstream steps of non-homologous end-joining (for example, XRCC4, DNA ligase 3 or DNA ligase 4) were identified, suggesting a NHEJ-independent function for the members of the ELF5 complex.

Some studies have suggested that the association of DNA repair proteins and transcription factors is opportunistic. A recent screen, for example, found that more than 70% of tested transcription factors (with no known involvement in the DNA damage response) were rapidly recruited to sites of laser-induced DNA damage and were reliant on PARP activity for this localisation. The study hypothesised that PARP1 induces chromatin remodelling (either directly or indirectly) to facilitate DNA repair, which also increases the accessibility of transcription factor DNA binding sites. However, the functional effects of the increased transcription factor binding on gene expression were not tested (Izhar *et al.*, 2015). This study suggests that the binding of transcription factors to sites of DNA damage may be a widespread opportunistic occurrence that is not directly relevant to transcriptional regulation.

In contrast to this perspective, other studies have demonstrated that DNA repair proteins such as DNA-PKcs, PARP1 and TOP2B are essential for gene regulation by various transcription factors, particularly those associated with developmental or stimulus-induced gene expression (Brenner *et al.*, 2011; Foulds *et al.*, 2013; Goodwin *et al.*, 2015; Haffner *et al.*, 2010; Ju *et al.*, 2006; Medunjanin *et al.*, 2010b). DNA-PKcs,

for example, was identified as an essential member of an ETS transcriptional complex in prostate cancer cells. Approximately 50% of prostate cancers have a gene rearrangement that places an ETS factor such as ERG or ETV1 under the control of the androgen-regulated TMPRSS2 promoter and 5' UTR (ETS-positive cancers). The AR-driven increase in ERG expression leads to a transcriptional program that promotes cancer development (in association with other genomic alterations such as PTEN loss), proliferation and invasion (reviewed in Sizemore *et al.*, 2017). DNA-PKcs and PARP1 were identified as ERG-interacting proteins in ETS-positive prostate cancer cell lines and tissues and were recruited to genomic binding sites by ERG. Both DNA-PKcs and PARP1 were required for ERG-regulated gene expression, while knockdown of XRCC4 (required for one of the final steps of NHEJ) had no effect on ERG activity. Furthermore, depletion or inhibition of PARP1 decreased the growth of prostate cancer cell xenografts (Brenner *et al.*, 2011). Collectively, these results indicate a NHEJ-independent transcriptional role for the ERG-DNA-PKcs-PARP1 complex through which ETS-driven transcriptional programs could be pharmacologically targeted. The identification of DNA-PKcs, and possibly PARP1, as ELF5-interacting proteins suggests a similar mechanism could function in the subset of breast cancers in which ELF5 is overexpressed. This includes approximately one-third of basal-like breast cancers (TCGA RNA-seq and METABRIC microarray cohorts, z-score greater than 2.0) and, as suggested by pre-clinical studies, ER-positive breast cancers that are resistant to endocrine therapy (Kalyuga *et al.*, 2012).

The effect of DNA-PKcs on ELF5 transcriptional activity has not been clearly established by this project. Expression of ELF5-regulated genes was used as an indirect measure of ELF5 transcriptional activity in cell lines overexpressing ELF5 with or without siRNA-mediated DNA-PKcs depletion (Figures 5.18 and 5.19). Overall, however, the results of these experiments were unclear, with varying effects of DNA-PKcs knockdown dependent on the gene and cell line. In addition, both the T47D and MDA-MB-231 cell lines showed minimal effects of increased ELF5 on gene expression, making interpretation of changes in ELF5 activity resulting from DNA-PKcs knockdown difficult. One of the most striking effects of DNA-PKcs knockdown in the MCF7 cell line was the large increase in ELF5-induced *PIP*, *VTGN1* and *GRHL3* expression. This suggests an inhibitory effect of DNA-PKcs on ELF5 transcriptional activity, in contrast to the activating effect seen in ETS-positive prostate cancer cells described above. However, it is unclear if these are direct effects arising from changes in ELF5 activity or indirect effects arising, for example, from alterations in the activity of other DNA-PKcs-regulated transcription factors such as ER. Other limitations include the relatively small

gene panel that was assessed, and the robust phenotype arising from DNA-PKcs knockdown, which may obscure the effects on ELF5.

On examining the set of transcription factors known to interact with DNA-PKcs, the activation of transcriptional activity by DNA-PKcs appears to be much more common than inhibition (Table 1.8, Chapter 1). One exception to this may be ETS1, which is known to interact with both DNA-PKcs and Ku80. Increased Ku80 overexpression was shown to inhibit ETS1 transcriptional activity; however, overexpression of DNA-PKcs alone had no effect (Choul-li *et al.*, 2009). Further experiments directly measuring ELF5 transcriptional activity (for example, reporter assays) and global changes in gene expression (for example, microarrays or RNA-sequencing) would help to clarify the role of DNA-PKcs in the regulation of ELF5 transcriptional activity.

The mechanisms by which DNA-PKcs might regulate ELF5 are currently unknown. One strong possibility, as discussed above, is DNA-PKcs-mediated phosphorylation of ELF5. DNA-PKcs has also been shown to be important in the regulation of co-factor dynamics, recruiting and dismissing various co-factors in a phosphorylation-dependent manner (Foulds *et al.*, 2013; Jeyakumar *et al.*, 2007). In addition, DNA-PKcs also regulates ELF5 mRNA and protein expression (discussed further below). Finally, DNA-PKcs may indirectly modify ELF5 expression and activity through actions on other transcription factors, for example ER, PR or SNAIL1/2; the overall effects of DNA-PKcs on the transcriptional program may therefore be determined by the balance of transcription factors expressed in the cell.

### **Site of interaction between ELF5 and DNA-PKcs**

One possible site of interaction between ELF5 and DNA-PKcs is the Pointed domain, which is believed to function in protein-protein interactions. ELF5 isoform 3, which lacks the Pointed domain, was shown in Chapter 3 to be able to regulate ELF5 target genes in a very similar way to the full-length isoforms 1 and 2. If the interaction were mediated via the Pointed domain, the transcriptional competence of Isoform 3 would argue against an ELF5-activating function of DNA-PKcs. In addition, there does not seem to be evidence of a release of inhibition on the transcriptional activity of ELF5 Isoform 3 (as was seen in the case of DNA-PKcs knockdown) compared to Isoform 2. These results suggest that the site of interaction lies outside the Pointed domain.

Consistent with this hypothesis, the site of interaction between DNA-PKcs and several ETS proteins in prostate cancer cells was shown to be within the ETS domain. ERG

interacts with DNA-PKcs through a key residue in the ETS domain (Y373 of ERG, in the sequence context RALRYYYDKN), which is conserved in the other DNA-PKcs-interacting ETS factors (ETS1, ETV1, SPI1) as well as ELF5 (sequence context RALRYYYKTG) (Brenner *et al.*, 2011). Interestingly, this tyrosine residue is immediately adjacent to the two conserved arginine residues (**RALR**) that have shown to be essential for binding of ETS factors to DNA, although it is unknown whether this is functionally significant (Bosselut *et al.*, 1993). While DNA-PKcs was shown to regulate the transcriptional activity of ERG in this study, the phosphorylation status of ERG was not examined, and it is therefore unknown if phosphorylation by DNA-PKcs is part of the regulatory mechanism. The tyrosine residue essential to the ERG interaction cannot be phosphorylated by DNA-PKcs (which only targets serine and threonine residues), however it could represent a kinase docking site. ETS1, for example, contains a docking site for MAPK1, which subsequently phosphorylates distal residues (reviewed in Garrett-Sinha, 2013; Hollenhorst *et al.*, 2011).

### **Mutual regulation of DNA-PKcs and ELF5**

Another finding in this chapter is the mutual regulation of DNA-PKcs and ELF5 in a reciprocal regulatory loop. DNA-PKcs knockdown resulted in an increase in ELF5 mRNA expression in both MCF7- and T47D- ELF5 cell lines (reaching statistical significance in the T47D lines only), indicating that DNA-PKcs negatively regulates ELF5 mRNA expression. This occurred in the vehicle-treated cells (which could be related to an increase in endogenous ELF5 expression) but, interestingly, also occurred in the doxycycline-treated cells (which, at least to some extent, must be related to an increase in ELF5 expression from the exogenous vector). This suggests a mechanism by which DNA-PKcs can regulate ELF5 mRNA levels even when the ELF5 gene is not within its normal genomic regulatory context. One mechanism by which this could occur is through post-transcriptional regulation of mRNA levels. DNA-PKcs has been shown to post-transcriptionally regulate MYC mRNA through inhibitory phosphorylation of polyribonucleotide nucleotidyltransferase 1 (PNPT1), an RNA-binding protein that degrades MYC mRNA, and it is possible that a similar regulatory mechanism may operate for ELF5 (Yu *et al.*, 2012).

Consistent with the mRNA results, DNA-PKcs knockdown also produced a small increase in ELF5-V5 protein expression in doxycycline-treated MCF7 and T47D- ELF5 lines. DNA-PKcs is known to regulate proteasomal degradation of several transcription factors, which could contribute to this increase in ELF5 protein levels independently of the effects on mRNA levels. Interaction of ER with DNA-PKcs, for example, prevents

ER ubiquitination, leading to increased ER protein levels (Medunjanin *et al.*, 2010b). In the MDA-MB-231 lines, no increase in ELF5 mRNA or protein expression was observed with DNA-PKcs knockdown; in fact, there was a slight decrease in ELF5 protein levels, indicating that the regulation of ELF5 by DNA-PKcs may depend on the expression of additional cell-type-specific factors.

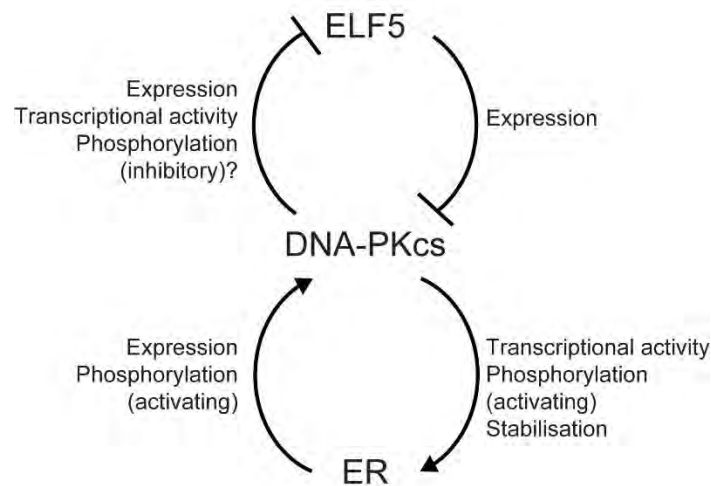
ELF5 also negatively regulates DNA-PKcs expression, demonstrated at the protein level for several T47D- and MDA-MB-231- ELF5 clonal cell lines. An ELF5 ChIP-seq was identified in the DNA-PKcs promoter, indicating that DNA-PKcs may be a direct ELF5 transcriptional target; however, DNA-PKcs mRNA expression was not altered in the MCF7-ELF5 RNA-sequencing experiments (Chapter 4). The DNA-PKcs promoter is actually a bidirectional promoter shared with the DNA replication licensing factor MCM4, which has been implicated in breast cancer development and progression (Kwok *et al.*, 2015). MCM4 expression is highly correlated with DNA-PKcs in The Cancer Genome Atlas cohort ( $r=0.83$ , Pearson's correlation), suggesting that these two factors may be co-regulated.

The phosphorylation of DNA-PKcs, a surrogate marker for activation, was also examined in ELF5-overexpressing cell lines. These experiments indicated that there was no change in DNA-PKcs S2056 phosphorylation in any cell line when ELF5 expression was induced. However, in contrast to previous reports, there was also no increase in S2056 phosphorylation when MCF7 cells were stimulated with oestradiol. This suggests that the experimental conditions used may have been sub-optimal for detecting subtle changes in phosphorylation.

### **A model for DNA-PKcs, ELF5 and ER transcriptional regulation**

Based on the above findings, a model for the interaction between DNA-PKcs, ELF5 and ER transcriptional regulation is proposed (Figure 5.24). DNA-PKcs and ER appear to operate in a classic positive feedback loop, with increases in the expression or activity of either protein positively regulating the other to amplify the response. In contrast, DNA-PKcs and ELF5 are proposed to operate in a mutual inhibitory loop, which, somewhat counterintuitively, also amplifies the response while simultaneously downregulating the ER loop. To illustrate, an increase in ELF5 expression in this model results in a decrease in DNA-PKcs expression. This decrease in DNA-PKcs, in turn, relieves the inhibition on ELF5, increasing ELF5 expression and (potentially) ELF5 transcriptional activity. At the same time, the reduction in DNA-PKcs levels leads to a decrease in ER protein levels and transcriptional activity, which feeds back onto DNA-

PKCs to further reduce its expression and activity. Thus, DNA-PKcs may mediate the balance between ER- and ELF5- directed cell fates.



**Figure 5.24: Model of ELF5, ER and DNA-PKcs interaction in breast cancer**

Based on published findings, DNA-PKcs and ER operate in a classic positive feedback loop, with increases in the expression or activity of either protein positively regulating the other to amplify the response. DNA-PKcs are proposed to operate in a mutual inhibitory loop; in this case, an alteration to the system also results in amplification of the response, while simultaneously downregulating the ER loop.

### Challenges in interpreting survival data

This chapter also presented the first subtype-specific analysis of the association between DNA-PKcs expression and survival. Unfortunately, no definite conclusions could be drawn due to discrepancies between different analyses, which may be related to several factors.

Firstly, different methods (RNA-sequencing and microarrays) have been used to measure expression, which are known to vary in their sensitivity of detection. However, even when using the same platform (for example, comparing Km-plotter and METABRIC cohorts) there are clear discrepancies in the analysis results, particularly in the basal-like subtype. One explanation for this may be the cut-offs used to separate low and high expression in these groups. The Km-plotter analysis, for example, used an automatically-generated cut-off (selected to give the best separation between groups), while the cBioPortal METABRIC analysis used a z-score cut-off of 2.0 to define upregulation. A z-score of 2.0 corresponds to 2 standard deviations above the mean expression for all breast cancer samples in the METABRIC cohort or



approximately the 98th percentile of expression (a fairly stringent but more objective cut-off). These cut-off settings, and the ability to manipulate them, are limited by the design of tools used (Km-plotter and cBioPortal). Due to these differences, the results generated by the two tools may not be directly comparable.

Another factor that makes interpretation of survival data difficult is that the measure of overall survival incorporates many deaths due to causes other than breast cancer. The use of overall survival as an endpoint is arguably less relevant for diseases such as breast cancer in the modern era. This is because advances in treatment, combined with earlier detection through screening, have dramatically improved prognosis to the point where approximately 2/3 of all diagnosed patients will live long enough to die from a cause other than breast cancer. This results in a significant dilution of deaths directly attributable to breast cancer and a loss of ability to reliably detect differences in outcomes (Cuzick, 2008; Cuzick, 2015).

Finally, DNA-PKcs has been shown to function in regulatory networks, both in DNA repair and transcription. Therefore, the expression and activity of associated proteins is likely to influence how DNA-PKcs functions in an individual tumour. One example is p53, with several studies suggesting that high DNA-PKcs in combination with wild-type p53 (compared to mutant p53) is associated with the best response to radiotherapy in breast and tonsillar carcinoma, although patient numbers in these studies were small (Friesland *et al.*, 2003; Soderlund Leifler *et al.*, 2010). Similarly, the transcriptional functions of DNA-PKcs that are emerging suggest a complex network of regulatory interactions. Therefore, the measurement of DNA-PKcs expression in isolation may not be an adequate marker for the functional effects of DNA-PKcs, even for breast cancers within the same molecular sub-group.

### **Therapeutic implications of the interaction between DNA-PKcs and ELF5**

The upregulation of DNA repair pathways in normal cells functions as a barrier to cancer progression. However, in cancer cells, upregulation of intact DNA repair mechanisms, combined with defects in apoptotic pathways, promotes cell survival and resistance to DNA-damaging therapies. Therefore, the inhibition of DNA repair may increase the sensitivity of cancer cells to DNA-damaging therapies, providing the rationale for the development of DNA repair inhibitors. As NHEJ is thought to be the major mechanism of repair for radiotherapy- and chemotherapy-induced double-stranded DNA breaks (DSBs), DNA-PKcs has been a major focus of these recent endeavours (Davidson *et al.*, 2013).

There are several small molecule inhibitors of DNA-PKcs, all of which inhibit (via various mechanisms) the catalytic kinase activity essential for the repair of DSBs. These include the non-specific PI3-kinase inhibitor wortmannin and the more selective LY294002 and related compounds (including NU7026 and NU7441) (reviewed in Cano and Harnor, 2017). Clinical development of small molecule DNA-PKcs inhibitors has been hampered by the short half-life of these compounds *in vivo*. However, promising results from a phase I clinical trial using the dual mTOR and DNA-PKcs inhibitor CC-115 in patients with chronic lymphocytic leukaemia (CLL) have recently been published, with 7 of 8 patients having a decrease in lymphadenopathy (Thijssen *et al.*, 2016). Three additional phase I trials using the inhibitors M3814 (also known as MSC2490484A) or VX-984 are ongoing, with the primary purpose of establishing pharmacokinetics and safety. These trials are using DNA-PK inhibitors in advanced solid tumours and/or CLL as a single agent (M3814, NCT02316197), in combination with radiotherapy and cisplatin (M3814, NCT02516813) or in combination with pegylated liposomal doxorubicin (VX-984, NCT02644278). Other approaches to clinical DNA-PKcs inhibition include antibody and nucleotide-based methods, for example the oligonucleotide-based treatment Dbait, which mimics DNA lesions and sequesters DNA repair proteins including DNA-PK and PARP1. Dbait is currently being tested in a phase I trial in cutaneous metastatic melanoma (NCT01469455).

However, it is now becoming clear that DNA-PKcs has many biological functions other than DNA repair, which may also be affected by DNA-PKcs inhibition. Understanding these additional functions is essential to the effective use of DNA-PKcs inhibitors in the clinical setting and may also assist in the screening of patients who are likely to benefit. Pre-clinical studies in prostate cancer, for example, indicate that patients with ETS-positive cancers may benefit from DNA-PKcs inhibition through a transcriptional mechanism that is distinct from the role of DNA-PKcs in DNA repair (Brenner *et al.*, 2011). Similarly, in ER-positive breast cancer, inhibition of DNA-PKcs could contribute to the downregulation of ER signalling leading to a reduction in tumour growth. However, in the model proposed above, the use of DNA-PKcs inhibitors in ER-positive breast cancers could also enhance ELF5 expression and transcriptional activity, potentially driving endocrine resistance and metastatic disease (Gallego-Ortega *et al.*, 2015; Kalyuga *et al.*, 2012). It is therefore imperative to understand how DNA-PKcs inhibition affects transcriptional networks, and not just DNA repair, in order to utilise these future treatments effectively.

### Unanswered questions about DNA-PKcs in transcriptional regulation

There are several unanswered questions about how DNA-PKcs functions in transcriptional regulation, which have potential clinical implications. Firstly, the mechanism of activation of DNA-PKcs in the absence of DNA damage is unknown. Recent studies have suggested that there may be Ku-independent mechanisms of DNA-PKcs activation; the autophosphorylation of DNA-PKcs in mitosis, for example, does not require expression of the Ku sub-units (Douglas *et al.*, 2014). Changes in the N-terminal conformation of DNA-PKcs have been shown to activate DNA-PKcs *in vitro* in a Ku-independent manner (Meek *et al.*, 2012). Therefore, one possibility is that DNA-PKcs is recruited to specific sites on undamaged DNA by transcription factors (rather than by the Ku sub-units) and that the interaction with the DNA-transcription factor complex leads to N-terminal conformational changes and enzymatic activation. This is one scenario for context-dependent activation of DNA-PKcs activity that could be explored by future studies.

A related question is whether the kinase activity of DNA-PKcs is required for the transcriptional functions of DNA-PKcs. The kinase activity is essential for the function of DNA-PKcs in DNA repair (Kurimasa *et al.*, 1999). However, the kinase domain of DNA-PKcs accounts for only 10% of the protein sequence or just over 25% if the stabilising FAT and FATC domains are included (Jette and Lees-Miller, 2015). Potential kinase-independent functions of other regions of DNA-PKcs have not been widely studied. Comparison of the effects of DNA-PKcs knockdown (using siRNA) or DNA-PKcs inhibition (using NU7441) in prostate cancer cells demonstrated that approximately half of all genes significantly up- or down-regulated by siRNA-mediated knockdown were not altered by kinase activity inhibition (Goodwin *et al.*, 2015). This could represent a set of genes uniquely regulated by kinase-independent functions of DNA-PKcs. A novel bromodomain (BRD)-like module has recently been identified in DNA-PKcs, which has been shown to interact with acetylated lysine 5 on H2AX resulting in activation of DNA-PKcs kinase activity (Wang *et al.*, 2015b). This raises some intriguing questions about what other histone modifications this module could recognise and whether these interactions might play a role in the transcriptional functions of DNA-PKcs. There is clearly much still to learn about how regions other than the kinase domain contribute to DNA-PKcs functions.

The clinical implications of the above questions relate to which functions of DNA-PKcs are being targeted by the current inhibitors, which suppress the kinase activity of DNA-PKcs. Although in some cases it might be beneficial to target both the DNA repair and

transcriptional functions of DNA-PKcs (for example, ETS-positive prostate cancer), in other situations this co-targeting could be potentially damaging (for example, in the ELF5-activating scenario for ER-positive breast cancer described above). It is essential to understand what functions of DNA-PKcs are being affected by kinase inhibition and, secondly, whether the individual functions of DNA-PKcs can be uniquely targeted. Potential routes for specific targeting of the DNA repair functions of DNA-PKcs include inhibition of the BRD-like module through the bromodomain and extra terminal (BET) inhibitor JQ1, which has been shown to bind to DNA-PKcs, or the targeting of scaffold molecules such as the recently identified lncRNA in NHEJ pathway 1 (LINP1) (Wang *et al.*, 2015b; Zhang *et al.*, 2016b). However, although no other functions have yet been described, the specificity of these mechanisms to DNA repair remains to be established.

## Chapter 6: Conclusions and Future Directions

Understanding the mechanisms by which transcription factors regulate gene expression, and in turn how transcription factors are themselves regulated, is an essential step in the development of transcription factor-targeted cancer therapies. The overall aim of this thesis was to investigate how the lineage-defining transcription factor ELF5 functions in breast cancer, and what additional factors regulate these functions.

At the most fundamental level, transcription factors function by binding to regulatory elements on DNA to alter expression of the associated gene. However, as discussed in Chapter 1, many additional factors influence this process, ranging from intrinsic properties of the transcription factor itself to global regulatory mechanisms such as chromatin accessibility.

The aspects of ELF5 function and regulation that have been explored in this thesis include:

- The expression of ELF5 in normal tissues and how this is altered in cancer;
- The expression and functions of ELF5 isoforms;
- The transcriptional targets of ELF5 in luminal breast cancer cells;
- The interplay between ELF5 and other transcription factors (FOXA1 and ER);
- The direct interactions of ELF5 with other transcription factors and co-factors, and
- The post-translational modifications of ELF5.

These findings have potential clinical implications and provide exciting new directions for future research.

In Chapter 3, the expression pattern and functions of ELF5 at the isoform level were investigated, demonstrating significantly altered expression in cancer. These alterations may drive abnormal cell fate decisions, suggesting that ELF5, like other ETS factors, may be a significant contributor to tumourigenesis. However, ELF5 cannot be clearly defined as either a tumour suppressor or an oncogene, as its expression, and most likely function, appear to be highly context-dependent. In kidney carcinoma, for example, the expression of ELF5 is almost universally down-regulated, suggesting a tumour suppressor function, whereas other tissues show variable up-regulation. The context-dependent expression and transcriptional effects may be related to co-operating oncogenic events or intrinsic features of the cell-of-origin (for example, the

chromatin state or the complement of other transcription factors and co-factors expressed in the cell). One question not addressed in this study was the mechanisms driving these alterations in ELF5 isoform expression in cancer. These may include changes in DNA methylation or altered upstream signalling pathways, providing avenues for future research into ELF5 dysregulation in cancer. The analysis in this chapter also provided a confident basis for the use of Isoform 2 in subsequent ELF5 over-expression studies, as this was shown to be the primary isoform expressed in breast cancer.

Due to the known roles of ELF5 in guiding the development of ER-negative mammary epithelial cells, an important goal of this research was to identify mechanisms by which increased ELF5 expression may promote the development of an oestrogen-insensitive breast cancer phenotype. In Chapter 4, next-generation sequencing technology provided a global overview of the transcriptional targets of ELF5 in the ER-positive luminal breast cancer cell line MCF7, leading to a number of new insights into ELF5 function. The ELF5 transcriptional signatures identified included the long-term adaptation to oestrogen-deprived conditions, MYC activation, and suppression of the interferon response, all of which could contribute to oestrogen-independent growth. An important future study will be the long-term over-expression of ELF5 in breast cancer cells, with monitoring of cell growth and gene expression profiles. This will provide insights into how breast cancer cells are able to overcome the initial growth-suppressive effects of ELF5 and move towards a proliferative, oestrogen-independent phenotype. Based on the findings presented here, this may involve the up-regulation of growth factor signalling pathways, similar to the adaptation of cells to LTED, and the action of proteins such as PIP and MYC. The cellular adaptation to long-term ELF5 expression may also alter the response to anti-oestrogen therapies, and cell line studies to determine if ELF5 expression alone can drive anti-oestrogen resistance may provide valuable insights into both *de novo* and acquired resistance. In addition, it will be interesting to determine if increased ELF5 expression can accelerate the process of LTED adaptation of breast cancer cells in culture.

Two additional areas of future research emerging from the transcriptomic studies are: (1) The relationship between ELF5 and MYC in normal alveolar development (which may provide new insights into their interaction in cancer), and (2) The role of ELF5 in the interferon response. ELF5 has recently been shown to enhance myeloid-derived suppressor cell infiltration and metastasis in breast cancer, and ELF5-driven suppression of IRF7 and the interferon response offers a compelling hypothesis for

how this might occur. Reactivation of the interferon signalling pathway in breast cancers with high ELF5 expression could decrease MDSC infiltration and metastasis. ELF5 expression may therefore have potential as a biomarker for response to interferon-based therapies. It will be important to test this hypothesis not only in cell culture but also in animal models, in order to study ELF5 and the interferon response in the context of an intact immune system and tumour microenvironment.

Another mechanism by which ELF5 may modulate the endocrine response is through the alteration of the binding sites of the ER pioneer factor FOXA1. The redistribution of FOXA1 binding is a phenomenon that has also been observed in tamoxifen-resistant cell lines and which likely underpins novel ER binding in poor-prognosis breast cancers. Therefore, this function of ELF5 could contribute to the “rewiring” of FOXA1 function in ER-positive breast cancers that progress and are ultimately lethal. Once more, this demonstrates that the function of a transcription factor is not solely determined by its intrinsic capabilities but also by direct and indirect interactions with other transcription factors. In this case, the interaction between ELF5 and FOXA1 is proposed to be indirect, for example through a “hit and run” mechanism, as FOXA1 was not identified as a direct ELF5 binding partner by RIME. This and other studies are now redefining the concept of a “pioneer” factor, suggesting that the ability of a transcription factor to mould the chromatin landscape and facilitate the binding of additional factors may depend on the context and co-factors, rather than an intrinsic property of the protein. However, the main limitation of this study was the lack of FOXA1 ChIP-seq replicates, especially given the relatively low sensitivity of the second replicate; this necessitated the use of strict criteria to define lost and gained FOXA1 binding sites. Further experimental replicates would help to more precisely define the subset of ELF5-driven FOXA1 binding sites and their functional implications. The analysis of how FOXA1 genomic binding changes over time could also be a valuable addition to the long-term ELF5 expression studies proposed above.

Finally, DNA-dependent protein kinase catalytic sub-unit (DNA-PKcs) was identified as an ELF5-interacting protein in human breast cancer cells. DNA-PKcs is emerging as an important regulator of various transcriptional networks in cancer. DNA-PKcs inhibitors have been developed to target the DNA repair functions of this protein, however the transcriptional functions also need to be considered if these are to be used in the clinical setting. An essential feature of the ELF5-DNA-PKcs interaction that needs to be addressed in future studies is the effect of DNA-PKcs on ELF5 transcriptional activity. The results presented in this thesis suggest a mutual inhibitory relationship, which has

potential implications for the use of DNA-PKcs inhibitors in the treatment of breast cancers with high ELF5 expression. However, the applicability of this model to global ELF5 transcriptional regulation is uncertain, as the observations are gene- and cell-type-specific. Future investigations to clarify this may include direct analysis of ELF5 transcriptional activity (for example, through reporter assays) and global analysis of gene expression following DNA-PKcs depletion in ELF5-expressing breast cancer cell lines. Another question is whether the interaction between ELF5 and DNA-PKcs occurs in basal-like breast cancers, in which ELF5 expression is frequently upregulated, and if the functional effects in this context are similar to those seen in luminal cell lines. The function of PARP1 in the ELF5-DNA-PKcs interaction is also an intriguing open question, which is particularly relevant given the relatively advanced state of PARP inhibitors in clinical trials.

Targeted mass spectrometry analysis of the ELF5 protein, in combination with phospho-site mutations, will help to define the role of phosphorylation in ELF5 function. In addition to DNA-PK, a number of other potential kinases were identified in the phospho-site analysis and RIME. The upstream signalling pathways leading to ELF5 phosphorylation are currently unknown, although growth factor and MAP kinase-mediated pathways are strong candidates for future investigation. Studies of ELF5 phosphorylation may also help to address the puzzling question of the role of the ELF5 Pointed domain. Threonine 93 in the Pointed domain is the highest-scoring predicted phospho-site in the ELF5 protein sequence and sits within a sequence matching the optimal DNA-PKcs motif. However, the gene expression studies in Chapter 3 indicate that the absence of the Pointed domain does not significantly impact ELF5 Isoform 3 transcriptional activity (at least in the context of over-expression). The specific identification of phosphorylated residues in the Pointed domain using mass spectrometry would point strongly towards a novel functional role for this domain.

Finally, there are a number of unanswered questions related to the transcriptional functions of DNA-PKcs. These include: (1) The mechanism of activation of DNA-PKcs in the absence of DNA damage (with one possibility being transcription factor-mediated conformational changes) and (2) Whether the kinase activity is required for the transcriptional functions of DNA-PKcs. An improved understanding of the activation and functional domains of DNA-PKcs could enable inhibitors to be developed that can selectively target these various biological functions. As in DNA repair, the dysregulation of DNA-PKcs in transcription may provide both therapeutic challenges and opportunities, and these will emerge from a detailed understanding of how DNA-PKcs



functions in transcriptional regulatory networks.

Transcriptional dysregulation has been described as “the most fundamental feature of cancer” (Gonda and Ramsay, 2015). Understanding the molecular basis for this dysregulation is therefore essential for the development of novel and targeted therapies. This thesis has provided new insights into the function and regulation of ELF5 in breast cancer, and represents an important contribution towards realising the potential of ELF5 as a therapeutic target in cancer.

## References

- Abdel-Fatah, T., Arora, A., Agarwal, D., Moseley, P., Perry, C., Thompson, N., *et al.* (2014). Adverse prognostic and predictive significance of low DNA-dependent protein kinase catalytic subunit (DNA-PKcs) expression in early-stage breast cancers. *Breast Cancer Res Treat*, 146(2), 309-320. doi: 10.1007/s10549-014-3035-2
- Abramson, J., Giraud, M., Benoist, C. and Mathis, D. (2010). Aire's Partners in the Molecular Control of Immunological Tolerance. *Cell*, 140(1), 123-135. doi: 10.1016/j.cell.2009.12.030
- Adam, R. C. and Fuchs, E. (2016). The Yin and Yang of Chromatin Dynamics In Stem Cell Fate Selection. *Trends Genet*, 32(2), 89-100. doi: 10.1016/j.tig.2015.11.002
- Adam, R. C., Yang, H., Rockowitz, S., Larsen, S. B., Nikolova, M., Oristian, D. S., *et al.* (2015). Pioneer factors govern super-enhancer dynamics in stem cell plasticity and lineage choice. *Nature*, 521(7552), 366-370. doi: 10.1038/nature14289
- Adelman, K. and Lis, J. T. (2012). Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans. *Nat Rev Genet*, 13(10), 720-731. doi: 10.1038/nrg3293
- Afgan, E., Baker, D., van den Beek, M., Blankenberg, D., Bouvier, D., Čech, M., *et al.* (2016). The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids Research*, 44(W1), W3-W10. doi: 10.1093/nar/gkw343
- Aguilar, H., Sole, X., Bonifaci, N., Serra-Musach, J., Islam, A., Lopez-Bigas, N., *et al.* (2010). Biological reprogramming in acquired resistance to endocrine therapy of breast cancer. *Oncogene*, 29(45), 6071-6083. doi: 10.1038/ncr.2010.333
- Ahn, B. Y., Elwi, A. N., Lee, B., Trinh, D. L. N., Klimowicz, A. C., Yau, A., *et al.* (2010). Genetic Screen Identifies Insulin-like Growth Factor Binding Protein 5 as a Modulator of Tamoxifen Resistance in Breast Cancer. *Cancer Research*, 70(8), 3013-3019. doi: 10.1158/0008-5472.can-09-3108
- Al-azawi, D., Ilroy, M. M., Kelly, G., Redmond, A. M., Bane, F. T., Cocchiola, S., *et al.* (2007). Ets-2 and p160 proteins collaborate to regulate c-Myc in endocrine resistant breast cancer. *Oncogene*, 27(21), 3021-3031.
- Allen, B. L. and Taatjes, D. J. (2015). The Mediator complex: a central integrator of transcription. *Nat Rev Mol Cell Biol*, 16(3), 155-166. doi: 10.1038/nrm3951
- Alles, M. C., Gardiner-Garden, M., Nott, D. J., Wang, Y., Foekens, J. A., Sutherland, R. L., *et al.* (2009). Meta-Analysis and Gene Set Enrichment Relative to ER Status Reveal Elevated Activity of MYC and E2F in the "Basal" Breast Cancer Subgroup. *PLoS One*, 4(3), e4710. doi: 10.1371/journal.pone.0004710
- Amatya, P. N., Kim, H.-B., Park, S.-J., Youn, C.-K., Hyun, J.-W., Chang, I.-Y., *et al.* (2012). A role of DNA-dependent protein kinase for the activation of AMP-activated protein kinase in response to glucose deprivation. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research*, 1823(12), 2099-2108. doi: 10.1016/j.bbamcr.2012.08.022
- American Type Culture Collection. (2015). *Breast Cancer Resource Book*. Manassas, Virginia, USA: ATCC.
- An, J., Xu, Q.-Z., Sui, J.-L., Bai, B. and Zhou, P.-K. (2005). Downregulation of c-myc protein by siRNA-mediated silencing of DNA-PKcs in HeLa cells. *International Journal of Cancer*, 117(4), 531-537. doi: 10.1002/ijc.21093
- An, J., Yang, D. Y., Xu, Q. Z., Zhang, S. M., Huo, Y. Y., Shang, Z. F., *et al.* (2008). DNA-dependent protein kinase catalytic subunit modulates the stability of c-Myc oncoprotein. *Mol Cancer*, 7, 32. doi: 10.1186/1476-4598-7-32
- Anbalagan, M. and Rowan, B. G. (2015). Estrogen receptor alpha phosphorylation and its functional impact in human breast cancer. *Mol Cell Endocrinol*, 418 Pt 3, 264-272. doi: 10.1016/j.mce.2015.01.016
- Anderson, C. W. and Appella, E. (2010). Signaling to the p53 Tumor Suppressor through Pathways Activated by Genotoxic and Non-Genotoxic Stresses. In E. A. Dennis & R. A. Bradshaw (Eds.), *Handbook of Cell Signaling (Second Edition)* (pp. 2185-2204). San Diego: Academic Press.
- Andres, A. C., van der Valk, M. A., Schonenberger, C. A., Fluckiger, F., LeMeur, M., Gerlinger, P., *et al.* (1988). Ha-ras and c-myc oncogene expression interferes with morphological and functional differentiation of mammary epithelial cells in single and double transgenic mice. *Genes Dev*, 2(11), 1486-1495.
- Araud, T., Genolet, R., Jaquier-Gubler, P. and Curran, J. (2007). Alternatively spliced isoforms of the human elk-1 mRNA within the 5' UTR: implications for ELK-1 expression. *Nucleic Acids Res*, 35(14), 4649-4663. doi: 10.1093/nar/gkm482
- Arinobu, Y., Mizuno, S., Chong, Y., Shigematsu, H., Iino, T., Iwasaki, H., *et al.* (2007). Reciprocal activation of GATA-1 and PU.1 marks initial specification of hematopoietic stem cells into myeloid and myelolymphoid lineages. *Cell Stem Cell*, 1(4), 416-427. doi: 10.1016/j.stem.2007.07.004
- Auckley, D. H., Crowell, R. E., Heaphy, E. R., Stidley, C. A., Lechner, J. F., Gilliland, F. D., *et al.* (2001). Reduced DNA-dependent protein kinase activity is associated with lung cancer. *Carcinogenesis*, 22(5), 723-727. doi: 10.1093/carcin/22.5.723
- Australian Institute of Health and Welfare. (2017). *Cancer in Australia 2017 Cancer series no.101* (pp. 204). Canberra: AIHW.
- Bach, L. A. (2015). Insulin-like growth factor binding proteins 4-6. *Best Practice & Research Clinical Endocrinology & Metabolism*, 29(5), 713-722. doi: 10.1016/j.beem.2015.06.002
- Bailey, S. M., Cornforth, M. N., Ullrich, R. L. and Goodwin, E. H. (2004). Dysfunctional mammalian telomeres join with DNA double-strand breaks. *DNA Repair (Amst)*, 3(4), 349-357. doi: 10.1016/j.dnarep.2003.11.007
- Bailey, S. M., Meyne, J., Chen, D. J., Kurimasa, A., Li, G. C., Lehnert, B. E., *et al.* (1999). DNA double-strand break repair proteins are required to cap the ends of mammalian chromosomes. *Proceedings of the National Academy of Sciences*, 96(26), 14899-14904. doi: 10.1073/pnas.96.26.14899
- Bailat, D., Begue, A., Stehelin, D. and Aumercier, M. (2002). ETS-1 transcription factor binds cooperatively to the palindromic head to head ETS-binding sites of the stromelysin-1 promoter by counteracting autoinhibition. *J Biol Chem*, 277(33), 29386-29398. doi: 10.1074/jbc.M200088200
- Ballare, C., Castellano, G., Gaveglia, L., Althammer, S., Gonzalez-Vallinas, J., Eyra, E., *et al.* (2013). Nucleosome-driven transcription factor binding and gene regulation. *Mol Cell*, 49(1), 67-79. doi: 10.1016/j.molcel.2012.10.019
- Banerji, J., Rusconi, S. and Schaffner, W. (1981). Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. *Cell*, 27(2 Pt 1), 299-308.
- Baniwal, S. K., Ching, N. O., Jordan, V. C., Tripathy, D. and Frenkel, B. (2014). Prolactin-induced protein (PIP)

- regulates proliferation of luminal A type breast cancer cells in an estrogen-independent manner. *PLoS One*, 8(6), e62361. doi: 10.1371/journal.pone.0062361
- Bannister, A. J., Gottlieb, T. M., Kouzarides, T. and Jackson, S. P. (1993). c-Jun is phosphorylated by the DNA-dependent protein kinase in vitro; definition of the minimal kinase recognition motif. *Nucleic Acids Res*, 21(5), 1289-1295.
- Baranello, L., Levens, D., Gupta, A. and Kouzine, F. (2012). The importance of being supercoiled: How DNA mechanics regulate dynamic processes. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, 1819(7), 632-638. doi: 10.1016/j.bbagr.2011.12.007
- Barozzi, I., Simonatto, M., Bonifacio, S., Yang, L., Rohs, R., Ghisletti, S., et al. (2014). Coregulation of transcription factor binding and nucleosome occupancy through DNA features of mammalian enhancers. *Mol Cell*, 54(5), 844-857. doi: 10.1016/j.molcel.2014.04.006
- Barrero, M. J. and Malik, S. (2013). The RNA polymerase II transcriptional machinery and its epigenetic context. *Subcell Biochem*, 61, 237-259. doi: 10.1007/978-94-007-4525-4\_11
- Bartkova, J., Horejsi, Z., Koed, K., Kramer, A., Tort, F., Zieger, K., et al. (2005). DNA damage response as a candidate anti-cancer barrier in early human tumorigenesis. *Nature*, 434(7035), 864-870. doi: 10.1038/nature03482
- Bates, D. O., Cui, T. G., Doughty, J. M., Winkler, M., Sugiono, M., Shields, J. D., et al. (2002). VEGF165b, an inhibitory splice variant of vascular endothelial growth factor, is down-regulated in renal cell carcinoma. *Cancer Res*, 62(14), 4123-4131.
- Becker, M. A., Hou, X., Harrington, S. C., Weroha, S. J., Gonzalez, S. E., Jacob, K. A., et al. (2012). IGFBP ratio confers resistance to IGF targeting and correlates with increased invasion and poor outcome in breast tumors. *Clin Cancer Res*, 18(6), 1808-1817. doi: 10.1158/1078-0432.CCR-11-1806
- Bell, A. C. and Felsenfeld, G. (2000). Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene. *Nature*, 405(6785), 482-485.
- Bell, O., Tiwari, V. K., Thoma, N. H. and Schubeler, D. (2011). Determinants and dynamics of genome accessibility. *Nat Rev Genet*, 12(8), 554-564. doi: 10.1038/nrg3017
- Benayoun, B. A. and Veitia, R. A. (2009). A post-translational modification code for transcription factors: sorting through a sea of signals. *Trends Cell Biol*, 19(5), 189-197. doi: 10.1016/j.tcb.2009.02.003
- Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1), 289-300.
- Bernard, V., Young, J., Chanson, P. and Binart, N. (2015). New insights in prolactin: pathological implications. *Nat Rev Endocrinol*, 11(5), 265-275. doi: 10.1038/nrendo.2015.36
- Bernardo, G. M., Bebek, G., Ginther, C. L., Sizemore, S. T., Lozada, K. L., Miedler, J. D., et al. (2012). FOXA1 represses the molecular phenotype of basal breast cancer cells. *Oncogene*. doi: 10.1038/onc.2012.62
- Bernardo, G. M. and Keri, R. A. (2012). FOXA1: a transcription factor with parallel functions in development and cancer. *Biosci Rep*, 32(2), 113-130. doi: 10.1042/BSR20110046
- Bernardo, G. M., Lozada, K. L., Miedler, J. D., Harburg, G., Hewitt, S. C., Mosley, J. D., et al. (2010). FOXA1 is an essential determinant of ERalpha expression and mammary ductal morphogenesis. *Development*, 137(12), 2045-2054. doi: 10.1242/dev.043299
- Beskow, C., Kanter, L., Holgersson, A., Nilsson, B., Frankendal, B., Avall-Lundqvist, E., et al. (2006). Expression of DNA damage response proteins and complete remission after radiotherapy of stage IB-IIA of cervical cancer. *Br J Cancer*, 94(11), 1683-1689. doi: 10.1038/sj.bjc.6603153
- Beskow, C., Skikuniene, J., Holgersson, A., Nilsson, B., Lewensohn, R., Kanter, L., et al. (2009). Radioresistant cervical cancer shows upregulation of the NHEJ proteins DNA-PKcs, Ku70 and Ku86. *Br J Cancer*, 101(5), 816-821.
- Bianchini, G., Balko, J. M., Mayer, I. A., Sanders, M. E. and Gianni, L. (2016). Triple-negative breast cancer: challenges and opportunities of a heterogeneous disease. *Nat Rev Clin Oncol*, 13(11), 674-690. doi: 10.1038/nrclinonc.2016.66
- Bianconi, E., Piovesan, A., Facchin, F., Beraudi, A., Casadei, R., Frabetti, F., et al. (2013). An estimation of the number of cells in the human body. *Ann Hum Biol*, 40(6), 463-471. doi: 10.3109/03014460.2013.807878
- Biddie, S. C., John, S., Sabo, P. J., Thurman, R. E., Johnson, T. A., Schiltz, R. L., et al. (2011). Transcription factor AP1 potentiates chromatin accessibility and glucocorticoid receptor binding. *Mol Cell*, 43(1), 145-155. doi: 10.1016/j.molcel.2011.06.016
- Bidwell, B. N., Slaney, C. Y., Withana, N. P., Forster, S., Cao, Y., Loi, S., et al. (2012). Silencing of Irf7 pathways in breast cancer cells promotes bone metastasis through immune escape. *Nat Med*, 18(8), 1224-1231. doi: 10.1038/nm.2830
- Biggin, M. D. (2011). Animal transcription networks as highly connected, quantitative continua. *Dev Cell*, 21(4), 611-626. doi: 10.1016/j.devcel.2011.09.008
- Bild, A. H., Yao, G., Chang, J. T., Wang, Q., Potti, A., Chasse, D., et al. (2006). Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature*, 439(7074), 353-357. doi: 10.1038/nature04296
- Bird, A. P. (1980). DNA methylation and the frequency of CpG in animal DNA. *Nucleic Acids Research*, 8(7), 1499-1504. doi: 10.1093/nar/8.7.1499
- Blackledge, N. P., Rose, N. R. and Klose, R. J. (2015). Targeting Polycomb systems to regulate gene expression: modifications to a complex story. *Nat Rev Mol Cell Biol*, 16(11), 643-649. doi: 10.1038/nrm4067
- Blackledge, N. P., Zhou, J. C., Tolstorukov, M. Y., Farcas, A. M., Park, P. J. and Klose, R. J. (2010). CpG islands recruit a histone H3 lysine 36 demethylase. *Mol Cell*, 38(2), 179-190. doi: 10.1016/j.molcel.2010.04.009
- Blair, D. G. and Athanasiou, M. (2000). Ets and retroviruses - transduction and activation of members of the Ets oncogene family in viral oncogenesis. *Oncogene*, 19(55), 6472-6481. doi: 10.1038/sj.onc.1204046
- Blakely, C. M., Sintasath, L., D'Cruz, C. M., Hahn, K. T., Dugan, K. D., Belka, G. K., et al. (2005). Developmental stage determines the effects of MYC in the mammary epithelium. *Development*, 132(5), 1147-1160. doi: 10.1242/dev.01655
- Blattler, A. and Farnham, P. J. (2013). Cross-talk between site-specific transcription factors and DNA methylation states. *J Biol Chem*, 288(48), 34287-34294. doi: 10.1074/jbc.R113.512517
- Boettiger, A. N., Bintu, B., Moffitt, J. R., Wang, S., Beliveau, B. J., Fudenberg, G., et al. (2016). Super-resolution imaging reveals distinct chromatin folding for different epigenetic states. *Nature*, 529(7586), 418-422. doi: 10.1038/nature16496
- Bonasio, R. and Shiekhattar, R. (2014). Regulation of transcription by long noncoding RNAs. *Annu Rev Genet*, 48, 433-455. doi: 10.1146/annurev-genet-120213-092323

- Bose, R., Karthaus, W. R., Armenia, J., Abida, W., Iaquinta, P. J., Zhang, Z., *et al.* (2017). ERF mutations reveal a balance of ETS factors controlling prostate oncogenesis. *Nature*, 546(7660), 671-675. doi: 10.1038/nature22820
- Bosselut, R., Levin, J., Adjadj, E. and Ghysdael, J. (1993). A single amino-acid substitution in the Ets domain alters core DNA binding specificity of Ets1 to that of the related transcription factors Elf1 and E74. *Nucleic Acids Res*, 21(22), 5184-5191.
- Bouchaert, P., Guerif, S., Debais, C., Irani, J. and Fromont, G. (2012). DNA-PKcs expression predicts response to radiotherapy in prostate cancer. *Int J Radiat Oncol Biol Phys*, 84(5), 1179-1185. doi: 10.1016/j.ijrobp.2012.02.014
- Bouker, K. B., Skaar, T. C., Fernandez, D. R., O'Brien, K. A., Riggins, R. B., Cao, D., *et al.* (2004). Interferon regulatory factor-1 mediates the proapoptotic but not cell cycle arrest effects of the steroidal antiestrogen ICI 182,780 (faslodex, fulvestrant). *Cancer Res*, 64(11), 4030-4039. doi: 10.1158/0008-5472.CAN-03-3602
- Bouquet, F., Ousset, M., Biard, D., Fallone, F., Dauvillier, S., Frit, P., *et al.* (2011). A DNA-dependent stress response involving DNA-PK occurs in hypoxic cells and contributes to cellular adaptation to hypoxia. *J Cell Sci*, 124(11), 1943-1951. doi: 10.1242/jcs.078030
- Boutet, Stéphane C., Cheung, Tom H., Quach, Navaline L., Liu, L., Prescott, S. L., Edalati, A., *et al.* (2012). Alternative Polyadenylation Mediates MicroRNA Regulation of Muscle Stem Cell Function. *Cell Stem Cell*, 10(3), 327-336. doi: 10.1016/j.stem.2012.01.017
- Bowie, M. L., Dietze, E. C., Delrow, J., Bean, G. R., Troch, M. M., Marjoram, R. J., *et al.* (2004). Interferon-regulatory factor-1 is critical for tamoxifen-mediated apoptosis in human mammary epithelial cells. *Oncogene*, 23(54), 8743-8755. doi: 10.1038/sj.onc.1208120
- Bozulic, L. and Hemmings, B. A. (2009). PIKKing on PKB: regulation of PKB activity by phosphorylation. *Curr Opin Cell Biol*, 21(2), 256-261. doi: 10.1016/j.ceb.2009.02.002
- Bracken, C. P., Scott, H. S. and Goodall, G. J. (2016). A network-biology perspective of microRNA function and dysfunction in cancer. *Nat Rev Genet*, 17(12), 719-732. doi: 10.1038/nrg.2016.134
- Brennan, C. W., Verhaak, R. G., McKenna, A., Campos, B., Nounshmehr, H., Salama, S. R., *et al.* (2013). The somatic genomic landscape of glioblastoma. *Cell*, 155(2), 462-477. doi: 10.1016/j.cell.2013.09.034
- Brenner, C., Deplus, R., Didelot, C., Lorient, A., Vire, E., De Smet, C., *et al.* (2005). Myc represses transcription through recruitment of DNA methyltransferase corepressor. *EMBO J*, 24(2), 336-346. doi: 10.1038/sj.emboj.7600509
- Brenner, J. C., Ateeq, B., Li, Y., Yocum, A. K., Cao, Q., Asangani, I. A., *et al.* (2011). Mechanistic rationale for inhibition of poly(ADP-ribose) polymerase in ETS gene fusion-positive prostate cancer. *Cancer Cell*, 19(5), 664-678. doi: 10.1016/j.ccr.2011.04.010
- Bretones, G., Delgado, M. D. and Leon, J. (2015). Myc and cell cycle control. *Biochim Biophys Acta*, 1849(5), 506-516. doi: 10.1016/j.bbagr.2014.03.013
- Bucceri, A., Kapitza, K. and Thoma, F. (2006). Rapid accessibility of nucleosomal DNA in yeast on a second time scale. *EMBO J*, 25(13), 3123-3132. doi: 10.1038/sj.emboj.7601196
- Buchwalter, G., Hickey, M. M., Cromer, A., Selfors, L. M., Gunawardane, R. N., Frishman, J., *et al.* (2013). PDEF promotes luminal differentiation and acts as a survival factor for ER-positive breast cancer cells. *Cancer Cell*, 23(6), 753-767. doi: 10.1016/j.ccr.2013.04.026
- Bullock, M. (2016). FOXO factors and breast cancer: outfoxing endocrine resistance. *Endocr Relat Cancer*, 23(2), R113-130. doi: 10.1530/ERC-15-0461
- Bunch, H., Zheng, X., Burkholder, A., Dillon, S. T., Motola, S., Birrane, G., *et al.* (2014). TRIM28 regulates RNA polymerase II promoter-proximal pausing and pause release. *Nat Struct Mol Biol*, 21(10), 876-883. doi: 10.1038/nsmb.2878
- Buschbeck, M. and Hake, S. B. (2017). Variants of core histones and their roles in cell fate decisions, development and cancer. *Nat Rev Mol Cell Biol*. doi: 10.1038/nrm.2016.166
- Bustamante, C. D., Fiedel-Alon, A., Williamson, S., Nielsen, R., Hubisz, M. T., Glanowski, S., *et al.* (2005). Natural selection on protein-coding genes in the human genome. *Nature*, 437(7062), 1153-1157. doi: 10.1038/nature04240
- Caizzi, L., Ferrero, G., Cutrupi, S., Cordero, F., Ballare, C., Miano, V., *et al.* (2014). Genome-wide activity of unliganded estrogen receptor- $\alpha$  in breast cancer cells. *Proc Natl Acad Sci U S A*, 111(13), 4892-4897. doi: 10.1073/pnas.1315445111
- Calo, E. and Wysocka, J. (2013). Modification of enhancer chromatin: what, how, and why? *Mol Cell*, 49(5), 825-837. doi: 10.1016/j.molcel.2013.01.038
- Cancer Genome Atlas Research Network. (2008). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, 455(7216), 1061-1068. doi: 10.1038/nature07385
- Cancer Genome Atlas Research Network. (2011). Integrated genomic analyses of ovarian carcinoma. *Nature*, 474(7353), 609-615. doi: 10.1038/nature10166
- Cancer Genome Atlas Research Network. (2012a). Comprehensive genomic characterization of squamous cell lung cancers. *Nature*, 489(7417), 519-525. doi: 10.1038/nature11404
- Cancer Genome Atlas Research Network. (2012b). Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, 487(7407), 330-337. doi: 10.1038/nature11252
- Cancer Genome Atlas Research Network. (2012c). Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418), 61-70. doi: 10.1038/nature11412
- Cancer Genome Atlas Research Network. (2013a). Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature*, 499(7456), 43-49. doi: 10.1038/nature12222
- Cancer Genome Atlas Research Network. (2013b). Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *N Engl J Med*, 368(22), 2059-2074. doi: 10.1056/NEJMoa1301689
- Cancer Genome Atlas Research Network. (2014a). Comprehensive molecular characterization of urothelial bladder carcinoma. *Nature*, 507(7492), 315-322. doi: 10.1038/nature12965
- Cancer Genome Atlas Research Network. (2014b). Comprehensive molecular profiling of lung adenocarcinoma. *Nature*, 511(7511), 543-550. doi: 10.1038/nature13385
- Cancer Genome Atlas Research Network. (2014c). Integrated genomic characterization of papillary thyroid carcinoma. *Cell*, 159(3), 676-690. doi: 10.1016/j.cell.2014.09.050
- Cancer Genome Atlas Research Network. (2015a). Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature*, 517(7536), 576-582. doi: 10.1038/nature14129

- Cancer Genome Atlas Research Network. (2015b). Genomic Classification of Cutaneous Melanoma. *Cell*, 161(7), 1681-1696. doi: 10.1016/j.cell.2015.05.044
- Cancer Genome Atlas Research Network, Kandoth, C., Schultz, N., Cherniack, A. D., Akbani, R., Liu, Y., *et al.* (2013). Integrated genomic characterization of endometrial carcinoma. *Nature*, 497(7447), 67-73. doi: 10.1038/nature12113
- Cano, C. and Harnor, S. J. (2017). Targeting DNA-PK for Cancer Therapy. *ChemMedChem*, n/a-n/a. doi: 10.1002/cmdc.201700143
- Carroll, J. S., Liu, X. S., Brodsky, A. S., Li, W., Meyer, C. A., Szary, A. J., *et al.* (2005). Chromosome-wide mapping of estrogen receptor binding reveals long-range regulation requiring the forkhead protein FoxA1. *Cell*, 122(1), 33-43. doi: 10.1016/j.cell.2005.05.008
- Cerami, E., Gao, J., Dogrusoz, U., Gross, B. E., Sumer, S. O., Aksoy, B. A., *et al.* (2012). The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov*, 2(5), 401-404. doi: 10.1158/2159-8290.CD-12-0095
- Chakrabarti, R., Hwang, J., Andres Blanco, M., Wei, Y., Lukacisin, M., Romano, R. A., *et al.* (2012a). Elf5 inhibits the epithelial-mesenchymal transition in mammary gland development and breast cancer metastasis by transcriptionally repressing Snail2. *Nat Cell Biol*, 14(11), 1212-1222. doi: 10.1038/ncb2607
- Chakrabarti, R., Wei, Y., Romano, R. A., DeCoste, C., Kang, Y. and Sinha, S. (2012b). Elf5 regulates mammary gland stem/progenitor cell fate by influencing notch signaling. *Stem Cells*, 30(7), 1496-1508. doi: 10.1002/stem.1112
- Chan, C. M. W., Martin, L.-A., Johnston, S. R. D., Ali, S. and Dowsett, M. (2002). Molecular changes associated with the acquisition of oestrogen hypersensitivity in MCF-7 breast cancer cells on long-term oestrogen deprivation. *The Journal of Steroid Biochemistry and Molecular Biology*, 81(4), 333-341. doi: 10.1016/S0960-0760(02)00074-2
- Charlot, C., Dubois-Pot, H., Serchov, T., Tourrette, Y. and Wasylyk, B. (2010). A review of post-translational modifications and subcellular localization of Ets transcription factors: possible connection with cancer and involvement in the hypoxic response. *Methods Mol Biol*, 647, 3-30. doi: 10.1007/978-1-60761-738-9\_1
- Chatterjee, P., Choudhary, G. S., Alswillah, T., Xiong, X., Heston, W. D., Magi-Galluzzi, C., *et al.* (2015). The TMPRSS2-ERG Gene Fusion Blocks XRCC4-Mediated Nonhomologous End-Joining Repair and Radiosensitizes Prostate Cancer Cells to PARP Inhibition. *Mol Cancer Ther*, 14(8), 1896-1906. doi: 10.1158/1535-7163.mct-14-0865
- Chatterjee, R. and Vinson, C. (2012). CpG methylation recruits sequence specific transcription factors essential for tissue specific gene expression. *Biochim Biophys Acta*, 1819(7), 763-770. doi: 10.1016/j.bbaggm.2012.02.014
- Chen, E. Y., Tan, C. M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G. V., *et al.* (2013a). Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics*, 14, 128. doi: 10.1186/1471-2105-14-128
- Chen, J. and Weiss, W. A. (2015). Alternative splicing in cancer: implications for biology and therapy. *Oncogene*, 34(1), 1-14. doi: 10.1038/ncr.2013.570
- Chen, J. D. and Evans, R. M. (1995). A transcriptional co-repressor that interacts with nuclear hormone receptors. *Nature*, 377(6548), 454-457. doi: 10.1038/377454a0
- Chen, P., Zhao, J., Wang, Y., Wang, M., Long, H., Liang, D., *et al.* (2013b). H3.3 actively marks enhancers and primes gene transcription via opening higher-ordered chromatin. *Genes Dev*, 27(19), 2109-2124. doi: 10.1101/gad.222174.113
- Chen, Z., Wang, Y., Warden, C. and Chen, S. (2015). Cross-talk between ER and HER2 regulates c-MYC-mediated glutamine metabolism in aromatase inhibitor resistant breast cancer cells. *The Journal of Steroid Biochemistry and Molecular Biology*, 149, 118-127. doi: 10.1016/j.jsbmb.2015.02.004
- Cheng, L., Wang, P., Yang, S., Yang, Y., Zhang, Q., Zhang, W., *et al.* (2012). Identification of genes with a correlation between copy number and expression in gastric cancer. *BMC Med Genomics*, 5(1), 14. doi: 10.1186/1755-8794-5-14
- Cheng, X. (2014). Structural and functional coordination of DNA and histone methylation. *Cold Spring Harb Perspect Biol*, 6(8). doi: 10.1101/cshperspect.a018747
- Chia, S., Gradishar, W., Mauriac, L., Bines, J., Amant, F., Federico, M., *et al.* (2008). Double-Blind, Randomized Placebo Controlled Trial of Fulvestrant Compared With Exemestane After Prior Nonsteroidal Aromatase Inhibitor Therapy in Postmenopausal Women With Hormone Receptor-Positive, Advanced Breast Cancer: Results From EFECT. *Journal of Clinical Oncology*, 26(10), 1664-1670. doi: 10.1200/jco.2007.13.5822
- Chibazakura, T., Watanabe, F., Kitajima, S., Tsukada, K., Yasukochi, Y. and Teraoka, H. (1997). Phosphorylation of Human General Transcription Factors TATA-Binding Protein and Transcription Factor IIB by DNA-Dependent Protein Kinase. *Eur J Biochem*, 247(3), 1166-1173. doi: 10.1111/j.1432-1033.1997.01166.x
- Choi, H. J., Lui, A., Ogony, J., Jan, R., Sims, P. J. and Lewis-Wambi, J. (2015). Targeting interferon response genes sensitizes aromatase inhibitor resistant breast cancer cells to estrogen-induced cell death. *Breast Cancer Research*, 17(1), 6. doi: 10.1186/s13058-014-0506-7
- Choi, Y. S. and Sinha, S. (2006). Determination of the consensus DNA-binding sequence and a transcriptional activation domain for ESE-2. *Biochem J*, 398(3), 497-507. doi: 10.1042/BJ20060375
- Choul-li, S., Drobecq, H. and Aumercier, M. (2009). DNA-dependent protein kinase is a novel interaction partner for Ets-1 isoforms. *Biochem Biophys Res Commun*, 390(3), 839-844. doi: 10.1016/j.bbrc.2009.10.059
- Ciccia, A. and Elledge, S. J. (2010). The DNA Damage Response: Making It Safe to Play with Knives. *Molecular Cell*, 40(2), 179-204. doi: 10.1016/j.molcel.2010.09.019
- Cierpicki, T., Risner, L. E., Grembecka, J., Lukasik, S. M., Popovic, R., Omonkowska, M., *et al.* (2010). Structure of the MLL CXXC domain-DNA complex and its functional role in MLL-AF9 leukemia. *Nat Struct Mol Biol*, 17(1), 62-68. doi: 10.1038/nsmb.1714
- Ciriello, G., Gatza, M. L., Beck, A. H., Wilkerson, M. D., Rhie, S. K., Pastore, A., *et al.* (2015). Comprehensive Molecular Portraits of Invasive Lobular Breast Cancer. *Cell*, 163(2), 506-519. doi: 10.1016/j.cell.2015.09.033
- Cirillo, L. A., Lin, F. R., Cuesta, I., Friedman, D., Jarnik, M. and Zaret, K. S. (2002). Opening of compacted chromatin by early developmental transcription factors HNF3 (FoxA) and GATA-4. *Mol Cell*, 9(2), 279-289. doi: 10.1092/276502004598 [pii]
- Cirillo, L. A., McPherson, C. E., Bossard, P., Stevens, K., Cherian, S., Shim, E. Y., *et al.* (1998). Binding of the winged-helix transcription factor HNF3 to a linker histone site on the nucleosome. *EMBO J*, 17(1), 244-254. doi: 10.1093/emboj/17.1.244

- Clarke, R., Tyson, J. J. and Dixon, J. M. (2015). Endocrine resistance in breast cancer—An overview and update. *Mol Cell Endocrinol*, 418 Pt 3, 220-234. doi: 10.1016/j.mce.2015.09.035
- Collis, S. J., DeWeese, T. L., Jeggo, P. A. and Parker, A. R. (2005). The life and death of DNA-PK. *Oncogene*, 24(6), 949-961. doi: 10.1038/sj.onc.1208332
- Cooper, C. D., Newman, J. A. and Gileadi, O. (2014). Recent advances in the structural molecular biology of Ets transcription factors: interactions, interfaces and inhibition. *Biochem Soc Trans*, 42(1), 130-138. doi: 10.1042/BST20130227
- Cornell, L., Munck, J. M., Alsinet, C., Villanueva, A., Ogle, L., Willoughby, C. E., *et al.* (2015). DNA-PK—A Candidate Driver of Hepatocarcinogenesis and Tissue Biomarker That Predicts Response to Treatment and Survival. *Clinical Cancer Research*, 21(4), 925-933. doi: 10.1158/1078-0432.ccr-14-0842
- Cosgrove, M. S., Boeke, J. D. and Wolberger, C. (2004). Regulated nucleosome mobility and the histone code. *Nat Struct Mol Biol*, 11(11), 1037-1043. doi: 10.1038/nsmb851
- Cowley, D. O. and Graves, B. J. (2000). Phosphorylation represses Ets-1 DNA binding by reinforcing autoinhibition. *Genes Dev*, 14(3), 366-376.
- Cramer, P., Armache, K. J., Baumli, S., Benkert, S., Brueckner, F., Buchen, C., *et al.* (2008). Structure of eukaryotic RNA polymerases. *Annu Rev Biophys*, 37, 337-352. doi: 10.1146/annurev.biophys.37.032807.130008
- Cramer, P., Bushnell, D. A., Fu, J., Gnatt, A. L., Maier-Davis, B., Thompson, N. E., *et al.* (2000). Architecture of RNA polymerase II and implications for the transcription mechanism. *Science*, 288(5466), 640-649.
- Crocker, J., Noon, E. P. and Stern, D. L. (2016). The Soft Touch: Low-Affinity Transcription Factor Binding Sites in Development and Evolution. In P. M. Wassarman (Ed.), *Essays on Developmental Biology Part B, Current Topics in Developmental Biology* (Vol. 117, pp. 455-469). Retrieved from [www.ncbi.nlm.nih.gov/pubmed/26969995](http://www.ncbi.nlm.nih.gov/pubmed/26969995). doi: 10.1016/bs.ctdb.2015.11.018
- Cui, F., Fan, R., Chen, Q., He, Y., Song, M., Shang, Z., *et al.* (2015). The involvement of c-Myc in the DNA double-strand break repair via regulating radiation-induced phosphorylation of ATM and DNA-PKcs activity. *Mol Cell Biochem*, 406(1), 43-51. doi: 10.1007/s11010-015-2422-2
- Curtin, N. J. (2012). DNA repair dysregulation from cancer driver to therapeutic target. *Nat Rev Cancer*, 12(12), 801-817. doi: 10.1038/nrc3399
- Curtis, C., Shah, S. P., Chin, S. F., Turashvili, G., Rueda, O. M., Dunning, M. J., *et al.* (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486(7403), 346-352. doi: 10.1038/nature10983
- Cuzick, J. (2008). Primary endpoints for randomised trials of cancer therapy. *The Lancet*, 371(9631), 2156-2158. doi: 10.1016/S0140-6736(08)60933-2
- Cuzick, J. (2015). Statistical controversies in clinical research: long-term follow-up of clinical trials in cancer. *Ann Oncol*, 26(12), 2363-2366. doi: 10.1093/annonc/mdv392
- Dabydeen, S. A., Kang, K., Díaz-Cruz, E. S., Alamri, A., Axelrod, M. L., Bouker, K. B., *et al.* (2015). Comparison of tamoxifen and letrozole response in mammary preneoplasia of ER and aromatase overexpressing mice defines an immune-associated gene signature linked to tamoxifen resistance. *Carcinogenesis*, 36(1), 122-132. doi: 10.1093/carcin/bgu237
- Dantas Machado, A. C., Zhou, T., Rao, S., Goel, P., Rastogi, C., Lazarovici, A., *et al.* (2015). Evolving insights on how cytosine methylation affects protein–DNA binding. *Briefings in Functional Genomics*, 14(1), 61-73. doi: 10.1093/bfpg/elu040
- Davidson, D., Amrein, L., Panasci, L. and Aloyz, R. (2013). Small Molecules, Inhibitors of DNA-PK, Targeting DNA Repair, and Beyond. *Front Pharmacol*, 4, 5. doi: 10.3389/fphar.2013.00005
- Davis, C. F., Ricketts, C. J., Wang, M., Yang, L., Cherniack, A. D., Shen, H., *et al.* (2014). The somatic genomic landscape of chromophobe renal cell carcinoma. *Cancer Cell*, 26(3), 319-330. doi: 10.1016/j.ccr.2014.07.014
- de Klerk, E. and 't Hoen, P. A. C. (2015). Alternative mRNA transcription, processing, and translation: insights from RNA sequencing. *Trends in Genetics*, 31(3), 128-139. doi: 10.1016/j.tig.2015.01.001
- de la Rica, L., Rodríguez-Ubreva, J., García, M., Islam, A. B., Urquiza, J. M., Hernando, H., *et al.* (2013). PU.1 target genes undergo Tet2-coupled demethylation and DNMT3b-mediated methylation in monocyte-to-osteoclast differentiation. *Genome Biology*, 14(9), R99. doi: 10.1186/gb-2013-14-9-r99
- De Val, S., Chi, N. C., Meadows, S. M., Minovitsky, S., Anderson, J. P., Harris, I. S., *et al.* (2008). Combinatorial regulation of endothelial gene expression by ets and forkhead transcription factors. *Cell*, 135(6), 1053-1064. doi: 10.1016/j.cell.2008.10.049
- Dechassa, M. L. and Luger, K. (2011). Nucleosomes as Control Elements for Accessing the Genome *Genome Organization and Function in the Cell Nucleus* (pp. 55-87): Wiley-VCH Verlag GmbH & Co. KGaA.
- DeFazio, L. G., Stansel, R. M., Griffith, J. D. and Chu, G. (2002). Synapsis of DNA ends by DNA-dependent protein kinase. *EMBO J*, 21(12), 3192-3200. doi: 10.1093/emboj/cdf299
- Denko, N. C., Giaccia, A. J., Stringer, J. R. and Stambrook, P. J. (1994). The human Ha-ras oncogene induces genomic instability in murine fibroblasts within one cell cycle. *Proc Natl Acad Sci U S A*, 91(11), 5124-5128.
- Desjardins, G., Okon, M., Graves, B. J. and McIntosh, L. P. (2016). Conformational Dynamics and the Binding of Specific and Nonspecific DNA by the Autoinhibited Transcription Factor Ets-1. *Biochemistry*, 55(29), 4105-4118. doi: 10.1021/acs.biochem.6b00460
- Deweese, J. E. and Osheroff, N. (2009). The DNA cleavage reaction of topoisomerase II: wolf in sheep's clothing. *Nucleic Acids Research*, 37(3), 738-748. doi: 10.1093/nar/gkn937
- Dhalluin, C., Carlson, J. E., Zeng, L., He, C., Aggarwal, A. K. and Zhou, M. M. (1999). Structure and ligand of a histone acetyltransferase bromodomain. *Nature*, 399(6735), 491-496. doi: 10.1038/20974
- Di Cerbo, V., Mohn, F., Ryan, D. P., Montellier, E., Kacem, S., Tropberger, P., *et al.* (2014). Acetylation of histone H3 at lysine 64 regulates nucleosome dynamics and facilitates transcription. *Elife*, 3, e01632. doi: 10.7554/eLife.01632
- Dittmann, K., Mayer, C., Fehrenbacher, B., Schaller, M., Raju, U., Milas, L., *et al.* (2005). Radiation-induced epidermal growth factor receptor nuclear import is linked to activation of DNA-dependent protein kinase. *J Biol Chem*, 280(35), 31182-31189. doi: 10.1074/jbc.M506591200
- Djebali, S., Davis, C. A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., *et al.* (2012). Landscape of transcription in human cells. *Nature*, 489(7414), 101-108. doi: 10.1038/nature11233
- Do, P. M., Varanasi, L., Fan, S., Li, C., Kubacka, I., Newman, V., *et al.* (2012). Mutant p53 cooperates with ETS2 to promote etoposide resistance. *Genes Dev*, 26(8), 830-845. doi: 10.1101/gad.181685.111

- Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., *et al.* (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1), 15-21. doi: 10.1093/bioinformatics/bts635
- Domcke, S., Bardet, A. F., Adrian Ginno, P., Hartl, D., Burger, L. and Schubeler, D. (2015). Competition between DNA methylation and transcription factors determines binding of NRF1. *Nature*, 528(7583), 575-579. doi: 10.1038/nature16462
- Donnison, M., Beaton, A., Davey, H. W., Broadhurst, R., L'Huillier, P. and Pfeffer, P. L. (2005). Loss of the extraembryonic ectoderm in Elf5 mutants leads to defects in embryonic patterning. *Development*, 132(10), 2299-2308. doi: 10.1242/dev.01819
- Douglas, P., Ye, R., Trinkle-Mulcahy, L., Neal, J. A., De Wever, V., Morrice, N. A., *et al.* (2014). Polo-like kinase 1 (PLK1) and protein phosphatase 6 (PP6) regulate DNA-dependent protein kinase catalytic subunit (DNA-PKcs) phosphorylation in mitosis. *Biosci Rep*, 34(3). doi: 10.1042/BSR20140051
- Dukler, N., Gulko, B., Huang, Y. F. and Siepel, A. (2016). Is a super-enhancer greater than the sum of its parts? *Nat Genet*, 49(1), 2-3. doi: 10.1038/ng.3759
- Dunham, I., Kundaje, A., Aldred, S. F., Collins, P. J., Davis, C., Doyle, F., *et al.* (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414), 57-74. doi: 10.1038/nature11247
- Dutta, B., Pusztai, L., Qi, Y., Andre, F., Lazar, V., Bianchini, G., *et al.* (2012). A network-based, integrative study to identify core biological pathways that drive breast cancer clinical subtypes. *Br J Cancer*, 106(6), 1107-1116. doi: 10.1038/bjc.2011.584
- Dvinge, H., Kim, E., Abdel-Wahab, O. and Bradley, R. K. (2016). RNA splicing factors as oncoproteins and tumour suppressors. *Nat Rev Cancer*, 16(7), 413-430. doi: 10.1038/nrc.2016.51
- Dvir, A., Peterson, S. R., Knuth, M. W., Lu, H. and Dynan, W. S. (1992). Ku autoantigen is the regulatory component of a template-associated protein kinase that phosphorylates RNA polymerase II. *Proc Natl Acad Sci U S A*, 89(24), 11920-11924.
- Early Breast Cancer Trialists' Collaborative Group. (1998). Tamoxifen for early breast cancer: an overview of the randomised trials. *Lancet*, 351(9114), 1451-1467.
- Elliott, S. L., Crawford, C., Mulligan, E., Summerfield, G., Newton, P., Wallis, J., *et al.* (2011). Mitoxantrone in combination with an inhibitor of DNA-dependent protein kinase: a potential therapy for high risk B-cell chronic lymphocytic leukaemia. *British Journal of Haematology*, 152(1), 61-71. doi: 10.1111/j.1365-2141.2010.08425.x
- Enchev, R. I., Schulman, B. A. and Peter, M. (2015). Protein neddylation: beyond cullin-RING ligases. *Nat Rev Mol Cell Biol*, 16(1), 30-44. doi: 10.1038/nrm3919
- Erkizan, H. V., Kong, Y., Merchant, M., Schlottmann, S., Barber-Rotenberg, J. S., Yuan, L., *et al.* (2009). A small molecule blocking oncogenic protein EWS-FLI1 interaction with RNA helicase A inhibits growth of Ewing's sarcoma. *Nat Med*, 15(7), 750-756. doi: 10.1038/nm.1983
- Ernst, J. and Kellis, M. (2012). ChromHMM: automating chromatin-state discovery and characterization. *Nat Meth*, 9(3), 215-216. doi: 10.1038/nmeth.1906
- Espejel, S., Franco, S., Sgura, A., Gae, D., Bailey, S. M., Taccioli, G. E., *et al.* (2002). Functional interaction between DNA-PKcs and telomerase in telomere length maintenance. *EMBO J*, 21(22), 6275-6287.
- Espejel, S., Martin, M., Klatt, P., Martin-Caballero, J., Flores, J. M. and Blasco, M. A. (2004). Shorter telomeres, accelerated ageing and increased lymphoma in DNA-PKcs-deficient mice. *EMBO Rep*, 5(5), 503-509. doi: 10.1038/sj.embor.7400127
- Eswaran, J., Horvath, A., Godbole, S., Reddy, S. D., Mudvari, P., Ohshiro, K., *et al.* (2013). RNA sequencing of cancer reveals novel splicing alterations. *Sci Rep*, 3, 1689. doi: 10.1038/srep01689
- Evert, M., Frau, M., Tomasi, M. L., Latte, G., Simile, M. M., Seddaiu, M. A., *et al.* (2013). Deregulation of DNA-dependent protein kinase catalytic subunit contributes to human hepatocarcinogenesis development and has a putative prognostic value. *Br J Cancer*, 109(10), 2654-2664. doi: 10.1038/bjc.2013.606
- Fabre, K. M., Ramaiah, L., Dregalla, R. C., Desaintes, C., Weil, M. M., Bailey, S. M., *et al.* (2011). Murine Prkdc polymorphisms impact DNA-PKcs function. *Radiat Res*, 175(4), 493-500. doi: 10.1667/RR2431.1
- Farnham, P. J. (2009). Insights from genomic profiling of transcription factors. *Nat Rev Genet*, 10(9), 605-616. doi: 10.1038/nrg2636
- Fatica, A. and Bozzoni, I. (2014). Long non-coding RNAs: new players in cell differentiation and development. *Nat Rev Genet*, 15(1), 7-21. doi: 10.1038/nrg3606
- Feldmann, A., Ivanek, R., Murr, R., Gaidatzis, D., Burger, L. and Schubeler, D. (2013). Transcription factor occupancy can mediate active turnover of DNA methylation at regulatory regions. *PLoS Genet*, 9(12), e1003994. doi: 10.1371/journal.pgen.1003994
- Feng, F. Y., Brenner, J. C., Hussain, M. and Chinnaiyan, A. M. (2014). Molecular pathways: targeting ETS gene fusions in cancer. *Clin Cancer Res*, 20(17), 4442-4448. doi: 10.1158/1078-0432.CCR-13-0275
- Fenrick, R., Amann, J. M., Lutterbach, B., Wang, L., Westendorf, J. J., Downing, J. R., *et al.* (1999). Both TEL and AML-1 contribute repression domains to the t(12;21) fusion protein. *Mol Cell Biol*, 19(10), 6566-6574.
- Ferguson, B. J., Mansur, D. S., Peters, N. E., Ren, H. and Smith, G. L. (2012). DNA-PK is a DNA sensor for IRF-3-dependent innate immunity. *Elife*, 1, e00047. doi: 10.7554/eLife.00047
- Ferguson, D. O., Sekiguchi, J. M., Chang, S., Frank, K. M., Gao, Y., DePinho, R. A., *et al.* (2000). The nonhomologous end-joining pathway of DNA repair is required for genomic stability and the suppression of translocations. *Proceedings of the National Academy of Sciences*, 97(12), 6630-6633. doi: 10.1073/pnas.110152897
- Ferlay J, S. I., Ervik M, Dikshit R, Eser S, Mathers C, Rebelo M, Parkin DM, Forman D, Bray, F. (2013). GLOBOCAN 2012 v1.0, Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 11 [Internet]. Retrieved January, 2017
- Filhol, O., Giacosa, S., Wallez, Y. and Cochet, C. (2015). Protein kinase CK2 in breast cancer: the CK2 $\beta$  regulatory subunit takes center stage in epithelial plasticity. *Cellular and Molecular Life Sciences*, 72(17), 3305-3322. doi: 10.1007/s00018-015-1929-8
- Filion, G. J., Zhenilo, S., Salozhin, S., Yamada, D., Prokhortchouk, E. and Defossez, P. A. (2006). A family of human zinc finger proteins that bind methylated DNA and repress transcription. *Mol Cell Biol*, 26(1), 169-181. doi: 10.1128/MCB.26.1.169-181.2006
- Filtz, T. M., Vogel, W. K. and Leid, M. (2014). Regulation of transcription factor activity by interconnected post-translational modifications. *Trends Pharmacol Sci*, 35(2), 76-85. doi: 10.1016/j.tips.2013.11.005
- Findlay, V. J., LaRue, A. C., Turner, D. P., Watson, P. M. and Watson, D. K. (2013). Understanding the role of ETS-

- mediated gene regulation in complex biological processes. *Adv Cancer Res*, 119, 1-61. doi: 10.1016/B978-0-12-407190-2.00001-0
- Fleuren, E. D., Zhang, L., Wu, J. and Daly, R. J. (2016). The kinome 'at large' in cancer. *Nat Rev Cancer*, 16(2), 83-98. doi: 10.1038/nrc.2015.18
- Fog, C. K., Galli, G. G. and Lund, A. H. (2012). PRDM proteins: important players in differentiation and disease. *Bioessays*, 34(1), 50-60. doi: 10.1002/bies.201100107
- Fong, Y. W., Cattoglio, C. and Tjian, R. (2013). The intertwined roles of transcription and repair proteins. *Mol Cell*, 52(3), 291-302. doi: 10.1016/j.molcel.2013.10.018
- Foulds, C. E., Feng, Q., Ding, C., Bailey, S., Hunsaker, T. L., Malovannaya, A., et al. (2013). Proteomic analysis of coregulators bound to ERalpha on DNA and nucleosomes reveals coregulator dynamics. *Mol Cell*, 51(2), 185-199. doi: 10.1016/j.molcel.2013.06.007
- Francia, S., Michelini, F., Saxena, A., Tang, D., de Hoon, M., Anelli, V., et al. (2012). Site-specific DICER and DROSHA RNA products control the DNA-damage response. *Nature*, 488(7410), 231-235. doi: 10.1038/nature11179
- Frauer, C., Spada, F. and Leonhardt, H. (2011). DNA Methylation *Genome Organization and Function in the Cell Nucleus* (pp. 21-54): Wiley-VCH Verlag GmbH & Co. KGaA.
- Friesland, S., Kanter-Lewensohn, L., Tell, R., Munck-Wikland, E., Lewensohn, R. and Nilsson, A. (2003). Expression of Ku86 confers favorable outcome of tonsillar carcinoma treated with radiotherapy. *Head & Neck*, 25(4), 313-321. doi: 10.1002/hed.10199
- Fu, Y.-P., Yu, J.-C., Cheng, T.-C., Lou, M. A., Hsu, G.-C., Wu, C.-Y., et al. (2003). Breast Cancer Risk Associated with Genotypic Polymorphism of the Nonhomologous End-Joining Genes. *A Multigenic Study on Cancer Susceptibility*, 63(10), 2440-2446.
- Fuda, N. J., Ardehali, M. B. and Lis, J. T. (2009). Defining mechanisms that regulate RNA polymerase II transcription in vivo. *Nature*, 461(7261), 186-192. doi: 10.1038/nature08449
- Gabut, M., Samavarchi-Tehrani, P., Wang, X., Slobodeniuc, V., O'Hanlon, D., Sung, H.-K., et al. (2011). An Alternative Splicing Switch Regulates Embryonic Stem Cell Pluripotency and Reprogramming. *Cell*, 147(1), 132-146. doi: 10.1016/j.cell.2011.08.023
- Galang, C. K., Muller, W. J., Foos, G., Oshima, R. G. and Hauser, C. A. (2004). Changes in the expression of many Ets family transcription factors and of potential target genes in normal mammary tissue and tumors. *J Biol Chem*, 279(12), 11281-11292. doi: 10.1074/jbc.M311887200
- Gallego-Ortega, D., Ledger, A., Roden, D. L., Law, A. M., Magenau, A., Kikhtyak, Z., et al. (2015). ELF5 Drives Lung Metastasis in Luminal Breast Cancer through Recruitment of Gr1+ CD11b+ Myeloid-Derived Suppressor Cells. *PLoS Biol*, 13(12), e1002330. doi: 10.1371/journal.pbio.1002330
- Gallego-Ortega, D., Oakes, S. R., Lee, H. J., Piggan, C. L. and Ormandy, C. J. (2013). ELF5, normal mammary development and the heterogeneous phenotypes of breast cancer. *Breast Cancer Management*, 2(6), 489-498. doi: 10.2217/bmt.13.50
- Gama-Sosa, M. A., Slagel, V. A., Trewyn, R. W., Oxenhandler, R., Kuo, K. C., Gehrke, C. W., et al. (1983). The 5-methylcytosine content of DNA from human tumors. *Nucleic Acids Res*, 11(19), 6883-6894.
- Gamsjaeger, R., Webb, S. R., Lamonica, J. M., Billin, A., Blobel, G. A. and Mackay, J. P. (2011). Structural basis and specificity of acetylated transcription factor GATA1 recognition by BET family bromodomain protein Brd3. *Mol Cell Biol*, 31(13), 2632-2640. doi: 10.1128/MCB.05413-11
- Gao, J., Aksoy, B. A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S. O., et al. (2013). Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci Signal*, 6(269), p11. doi: 10.1126/scisignal.2004088
- Garrett-Sinha, L. A. (2013). Review of Ets1 structure, function, and roles in immunity. *Cell Mol Life Sci*, 70(18), 3375-3390. doi: 10.1007/s00018-012-1243-7
- Garvie, C. W., Hagman, J. and Wolberger, C. (2001). Structural studies of Ets-1/Pax5 complex formation on DNA. *Mol Cell*, 8(6), 1267-1276.
- Garvie, C. W., Pufall, M. A., Graves, B. J. and Wolberger, C. (2002). Structural analysis of the autoinhibition of Ets-1 and its role in protein partnerships. *J Biol Chem*, 277(47), 45529-45536. doi: 10.1074/jbc.M206327200
- Gerstein, M. B., Kundaje, A., Hariharan, M., Landt, S. G., Yan, K. K., Cheng, C., et al. (2012). Architecture of the human regulatory network derived from ENCODE data. *Nature*, 489(7414), 91-100. doi: 10.1038/nature11245
- Gertz, J., Savic, D., Varley, K. E., Partridge, E. C., Safi, A., Jain, P., et al. (2013). Distinct properties of cell-type-specific and shared transcription factor binding sites. *Mol Cell*, 52(1), 25-36. doi: 10.1016/j.molcel.2013.08.037
- Giffin, W., Kwast-Welfeld, J., Rodda, D. J., Prefontaine, G. G., Traykova-Andonova, M., Zhang, Y., et al. (1997). Sequence-specific DNA binding and transcription factor phosphorylation by Ku Autoantigen/DNA-dependent protein kinase. Phosphorylation of Ser-527 of the rat glucocorticoid receptor. *J Biol Chem*, 272(9), 5647-5658.
- Gilley, D., Tanaka, H., Hande, M. P., Kurimasa, A., Li, G. C., Oshimura, M., et al. (2001). DNA-PKcs is critical for telomere capping. *Proc Natl Acad Sci U S A*, 98(26), 15084-15088. doi: 10.1073/pnas.261574698
- Giraud, M., Yoshida, H., Abramson, J., Rahl, P. B., Young, R. A., Mathis, D., et al. (2012). Aire unleashes stalled RNA polymerase to induce ectopic gene expression in thymic epithelial cells. *Proc Natl Acad Sci U S A*, 109(2), 535-540. doi: 10.1073/pnas.1119351109
- Gonda, T. J. and Ramsay, R. G. (2015). Directly targeting transcriptional dysregulation in cancer. *Nat Rev Cancer*, 15(11), 686-694. doi: 10.1038/nrc4018
- Goodrich, J. A. and Tjian, R. (2010). Unexpected roles for core promoter recognition factors in cell-type-specific transcription and gene regulation. *Nat Rev Genet*, 11(8), 549-558. doi: 10.1038/nrg2847
- Goodwin, J. F. and Knudsen, K. E. (2014). Beyond DNA repair: DNA-PK function in cancer. *Cancer Discov*, 4(10), 1126-1139. doi: 10.1158/2159-8290.CD-14-0358
- Goodwin, J. F., Kothari, V., Drake, J. M., Zhao, S., Dylgjeri, E., Dean, J. L., et al. (2015). DNA-PKcs-Mediated Transcriptional Regulation Drives Prostate Cancer Progression and Metastasis. *Cancer Cell*, 28(1), 97-113. doi: 10.1016/j.ccell.2015.06.004
- Goodwin, J. F., Schiewer, M. J., Dean, J. L., Schrecengost, R. S., de Leeuw, R., Han, S., et al. (2013). A hormone-DNA repair circuit governs the response to genotoxic insult. *Cancer Discov*, 3(11), 1254-1271. doi: 10.1158/2159-8290.CD-13-0108
- Gorgoulis, V. G., Vassiliou, L.-V. F., Karakaidos, P., Zacharatos, P., Kotsinas, A., Liloglou, T., et al. (2005). Activation of the DNA damage checkpoint and genomic instability in human precancerous lesions. *Nature*, 434(7035), 907-913. doi: 10.1038/nature03485



- Gosline, S. J. C., Gurtan, A. M., JnBaptiste, C. K., Bosson, A., Milani, P., Dalin, S., *et al.* (2016). Elucidating MicroRNA Regulatory Networks Using Transcriptional, Post-transcriptional, and Histone Modification Measurements. *Cell Reports*, 14(2), 310-319. doi: 10.1016/j.celrep.2015.12.031
- Graf, T. and Enver, T. (2009). Forcing cells to change lineages. *Nature*, 462(7273), 587-594. doi: 10.1038/nature08533
- Gray, D. C., Hoeflich, K. P., Peng, L., Gu, Z., Gogineni, A., Murray, L. J., *et al.* (2007). pHUSH: a single vector system for conditional gene expression. *BMC Biotechnol*, 7, 61. doi: 10.1186/1472-6750-7-61
- Gray, K. A., Yates, B., Seal, R. L., Wright, M. W. and Bruford, E. A. (2015). Genenames.org: the HGNC resources in 2015. *Nucleic Acids Res*, 43(Database issue), D1079-1085. doi: 10.1093/nar/gku1071
- Green, A. R., Aleskandarany, M. A., Agarwal, D., Elsheikh, S., Nolan, C. C., Diez-Rodriguez, M., *et al.* (2016). MYC functions are specific in biological subtypes of breast cancer and confers resistance to endocrine therapy in luminal tumours. *Br J Cancer*, 114(8), 917-928. doi: 10.1038/bjc.2016.46
- Green, K. A. and Carroll, J. S. (2007). Oestrogen-receptor-mediated transcription and the influence of co-factors and chromatin state. *Nat Rev Cancer*, 7(9), 713-722. doi: 10.1038/nrc2211
- Gross, K., Wronski, A., Skibinski, A., Phillips, S. and Kuperwasser, C. (2016). Cell Fate Decisions During Breast Cancer Development. *J Dev Biol*, 4(1), 4. doi: 10.3390/jdb4010004
- Grossi, E., Sanchez, Y. and Huarte, M. (2016). Expanding the p53 regulatory network: LncRNAs take up the challenge. *Biochim Biophys Acta*, 1859(1), 200-208. doi: 10.1016/j.bbarm.2015.07.011
- GTEX Consortium. (2015). Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science*, 348(6235), 648-660. doi: 10.1126/science.1262110
- Guidez, F., Petrie, K., Ford, A. M., Lu, H., Bennett, C. A., MacGregor, A., *et al.* (2000). Recruitment of the nuclear receptor corepressor N-CoR by the TEL moiety of the childhood leukemia-associated TEL-AML1 oncoprotein. *Blood*, 96(7), 2557-2561.
- Guillouf, C., Gallais, I. and Moreau-Gachelin, F. (2006). Spi-1/PU.1 oncoprotein affects splicing decisions in a promoter binding-dependent manner. *J Biol Chem*, 281(28), 19145-19155. doi: 10.1074/jbc.M512049200
- Györfy, B., Lanczky, A., Eklund, A. C., Denkert, C., Budczies, J., Li, Q., *et al.* (2010). An online survival analysis tool to rapidly assess the effect of 22,277 genes on breast cancer prognosis using microarray data of 1,809 patients. *Breast Cancer Research and Treatment*, 123(3), 725-731. doi: 10.1007/s10549-009-0674-9
- Haffner, M. C., Aryee, M. J., Toubaji, A., Esopi, D. M., Albadine, R., Gurel, B., *et al.* (2010). Androgen-induced TOP2B-mediated double-strand breaks and prostate cancer gene rearrangements. *Nat Genet*, 42(8), 668-675. doi: 10.1038/ng.613
- Halazonetis, T. D., Gorgoulis, V. G. and Bartek, J. (2008). An Oncogene-Induced DNA Damage Model for Cancer Development. *Science*, 319(5868), 1352-1355. doi: 10.1126/science.1140735
- Hall, J. M., McDonnell, D. P. and Korach, K. S. (2002). Allosteric Regulation of Estrogen Receptor Structure, Function, and Coactivator Recruitment by Different Estrogen Response Elements. *Molecular Endocrinology*, 16(3), 469-486. doi: 10.1210/mend.16.3.0814
- Hall, M. A., Shundrovsky, A., Bai, L., Fulbright, R. M., Lis, J. T. and Wang, M. D. (2009). High-resolution dynamic mapping of histone-DNA interactions in a nucleosome. *Nat Struct Mol Biol*, 16(2), 124-129. doi: 10.1038/nsmb.1526
- Hammond-Martel, I., Yu, H. and Affar el, B. (2012). Roles of ubiquitin signaling in transcription regulation. *Cell Signal*, 24(2), 410-421. doi: 10.1016/j.cellsig.2011.10.009
- Han, B., Bhowmick, N., Qu, Y., Chung, S., Giuliano, A. E. and Cui, X. (2017). FOXC1: an emerging marker and therapeutic target for cancer. *Oncogene*, 36(28), 3957-3963. doi: 10.1038/onc.2017.48
- Hanahan, D. and Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. *Cell*, 144(5), 646-674. doi: 10.1016/j.cell.2011.02.013
- Hantsche, M. and Cramer, P. (2016). The Structural Basis of Transcription: 10 Years After the Nobel Prize in Chemistry. *Angew Chem Int Ed Engl*, 55(52), 15972-15981. doi: 10.1002/anie.201608066
- Harbeck, N. and Gnant, M. (2016). Breast cancer. *Lancet*. doi: 10.1016/S0140-6736(16)31891-8
- Harris, J., Stanford, P. M., Sutherland, K., Oakes, S. R., Naylor, M. J., Robertson, F. G., *et al.* (2006). Socs2 and elf5 mediate prolactin-induced mammary gland development. *Mol Endocrinol*, 20(5), 1177-1187. doi: 10.1210/me.2005-0473
- Harrow, J., Frankish, A., Gonzalez, J. M., Tapanari, E., Diekhans, M., Kokocinski, F., *et al.* (2012). GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res*, 22(9), 1760-1774. doi: 10.1101/gr.135350.111
- Hassler, M. R., Redl, E., Hudson, Q. J., Miller, W. J. and Egger, G. (2016). Chapter 1 - Basic Epigenetic Mechanisms and Phenomena *Drug Discovery in Cancer Epigenetics* (pp. 3-40). Boston: Academic Press.
- Hatano, H., Kudo, Y., Ogawa, I., Tsunematsu, T., Kikuchi, A., Abiko, Y., *et al.* (2008). IFN-induced transmembrane protein 1 promotes invasion at early stage of head and neck cancer progression. *Clin Cancer Res*, 14(19), 6097-6105. doi: 10.1158/1078-0432.CCR-07-4761
- Hay, D., Hughes, J. R., Babbs, C., Davies, J. O., Graham, B. J., Hanssen, L. L., *et al.* (2016). Genetic dissection of the alpha-globin super-enhancer in vivo. *Nat Genet*, 48(8), 895-903. doi: 10.1038/ng.3605
- He, J., Pan, Y., Hu, J., Albarracin, C., Wu, Y. and Dai, J. L. (2007). Profile of Ets gene expression in human breast carcinoma. *Cancer Biol Ther*, 6(1), 76-82.
- He, Y. F., Li, B. Z., Li, Z., Liu, P., Wang, Y., Tang, Q., *et al.* (2011). Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. *Science*, 333(6047), 1303-1307. doi: 10.1126/science.1210944
- Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., *et al.* (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell*, 38(4), 576-589. doi: 10.1016/j.molcel.2010.05.004
- Heinz, S., Romanoski, C. E., Benner, C. and Glass, C. K. (2015). The selection and function of cell type-specific enhancers. *Nat Rev Mol Cell Biol*, 16(3), 144-154. doi: 10.1038/nrm3949
- Heldring, N., Isaacs, G. D., Diehl, A. G., Sun, M., Cheung, E., Ranish, J. A., *et al.* (2011). Multiple sequence-specific DNA-binding proteins mediate estrogen receptor signaling through a tethering pathway. *Mol Endocrinol*, 25(4), 564-574. doi: 10.1210/me.2010-0425
- Hellems, J., Mortier, G., De Paepe, A., Speleman, F. and Vandesompele, J. (2007). qBase relative quantification framework and software for management and automated analysis of real-time quantitative PCR data. *Genome Biol*, 8(2), R19. doi: 10.1186/gb-2007-8-2-r19
- Hemmings, B. A. and Restuccia, D. F. (2012). PI3K-PKB/Akt pathway. *Cold Spring Harb Perspect Biol*, 4(9), a011189.

- doi: 10.1101/cshperspect.a011189
- Henikoff, S. and Shilatifard, A. (2011). Histone modification: cause or cog? *Trends Genet*, 27(10), 389-396. doi: 10.1016/j.tig.2011.06.006
- Hennighausen, L. and Robinson, G. W. (2005). Information networks in the mammary gland. *Nat Rev Mol Cell Biol*, 6(9), 715-725. doi: 10.1038/nrm1714
- Hergeth, S. P. and Schneider, R. (2015). The H1 linker histones: multifunctional proteins beyond the nucleosomal core particle. *EMBO Rep*, 16(11), 1439-1453. doi: 10.15252/embr.201540749
- Hermani, A., Shukla, A., Medunjanin, S., Werner, H. and Mayer, D. (2013). Insulin-like growth factor binding protein-4 and -5 modulate ligand-dependent estrogen receptor- $\alpha$  activation in breast cancer cells in an IGF-independent manner. *Cell Signal*, 25(6), 1395-1402. doi: 10.1016/j.cellsig.2013.02.018
- Herschkowitz, J. I., Simin, K., Weigman, V. J., Mikaelian, I., Usary, J., Hu, Z., et al. (2007). Identification of conserved gene expression features between murine mammary carcinoma models and human breast tumors. *Genome Biol*, 8(5), R76. doi: 10.1186/gb-2007-8-5-r76
- Hervouet, E., Vallette, F. M. and Cartron, P. F. (2009). Dnmt3/transcription factor interactions as crucial players in targeted DNA methylation. *Epigenetics*, 4(7), 487-499.
- Herz, H. M., Garruss, A. and Shilatifard, A. (2013). SET for life: biochemical activities and biological functions of SET domain-containing proteins. *Trends Biochem Sci*, 38(12), 621-639. doi: 10.1016/j.tibs.2013.09.004
- Hilton, H. N., Kalyuga, M., Cowley, M. J., Alles, M. C., Lee, H. J., Caldon, C. E., et al. (2010). The antiproliferative effects of progestins in T47D breast cancer cells are tempered by progestin induction of the ETS transcription factor Elf5. *Mol Endocrinol*, 24(7), 1380-1392. doi: 10.1210/me.2009-0516
- Hiscox, S. G., Julia Gee, Nicholson, Robert I. (2009). *Therapeutic Resistance to Anti-Hormonal Drugs in Breast Cancer* doi:10.1007/978-1-4020-8526-0
- Hnisz, D., Abraham, B. J., Lee, T. I., Lau, A., Saint-Andre, V., Sigova, A. A., et al. (2013). Super-enhancers in the control of cell identity and disease. *Cell*, 155(4), 934-947. doi: 10.1016/j.cell.2013.09.053
- Hock, A. K. and Vousden, K. H. (2014). The role of ubiquitin modification in the regulation of p53. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research*, 1843(1), 137-149. doi: 10.1016/j.bbamcr.2013.05.022
- Hoeijmakers, J. H. J. (2009). DNA Damage, Aging, and Cancer. *New England Journal of Medicine*, 361(15), 1475-1485. doi: 10.1056/NEJMra0804615
- Holgerson, A., Nilsson, A., Lewensohn, R. and Kanter, L. (2004). Expression of DNA-PKcs and Ku86, but not Ku70, differs between lymphoid malignancies. *Experimental and Molecular Pathology*, 77(1), 1-6. doi: 10.1016/j.yexmp.2004.02.001
- Hollenhorst, P. C., Chandler, K. J., Poulsen, R. L., Johnson, W. E., Speck, N. A. and Graves, B. J. (2009). DNA specificity determinants associate with distinct transcription factor functions. *PLoS Genet*, 5(12), e1000778. doi: 10.1371/journal.pgen.1000778
- Hollenhorst, P. C., McIntosh, L. P. and Graves, B. J. (2011). Genomic and biochemical insights into the specificity of ETS transcription factors. *Annu Rev Biochem*, 80, 437-471. doi: 10.1146/annurev.biochem.79.081507.103945
- Holmberg, C. I., Tran, S. E., Eriksson, J. E. and Sistonen, L. (2002). Multisite phosphorylation provides sophisticated regulation of transcription factors. *Trends Biochem Sci*, 27(12), 619-627.
- Holmes, K. A., Hurtado, A., Brown, G. D., Launchbury, R., Ross-Innes, C. S., Hadfield, J., et al. (2012). Transducin-like enhancer protein 1 mediates estrogen receptor binding and transcriptional activity in breast cancer cells. *Proc Natl Acad Sci U S A*, 109(8), 2748-2753. doi: 10.1073/pnas.1018863108
- Hoppe, P. S., Schwarzfischer, M., Loeffler, D., Kokkaliaris, K. D., Hilsenbeck, O., Moritz, N., et al. (2016). Early myeloid lineage choice is not initiated by random PU.1 to GATA1 protein ratios. *Nature*, 535(7611), 299-302. doi: 10.1038/nature18320
- Horn, S., Figl, A., Rachakonda, P. S., Fischer, C., Sucker, A., Gast, A., et al. (2013). TERT promoter mutations in familial and sporadic melanoma. *Science*, 339(6122), 959-961. doi: 10.1126/science.1230062
- Houtkooper, R. H., Pirinen, E. and Auwerx, J. (2012). Sirtuins as regulators of metabolism and healthspan. *Nat Rev Mol Cell Biol*, 13(4), 225-238. doi: 10.1038/nrm3293
- Hsieh, C. L., Fei, T., Chen, Y., Li, T., Gao, Y., Wang, X., et al. (2014). Enhancer RNAs participate in androgen receptor-driven looping that selectively enhances gene activation. *Proc Natl Acad Sci U S A*, 111(20), 7319-7324. doi: 10.1073/pnas.1324151111
- Hu, H., Gu, Y., Qian, Y., Hu, B., Zhu, C., Wang, G., et al. (2014). DNA-PKcs is important for Akt activation and gemcitabine resistance in PANC-1 pancreatic cancer cells. *Biochemical and Biophysical Research Communications*, 452(1), 106-111. doi: 10.1016/j.bbrc.2014.08.059
- Hu, S., Wan, J., Su, Y., Song, Q., Zeng, Y., Nguyen, H. N., et al. (2013). DNA methylation presents distinct binding sites for human transcription factors. *Elife*, 2, e00726. doi: 10.7554/eLife.00726
- Hu, Z., Gu, X., Baraoidan, K., Ibanez, V., Sharma, A., Kadkol, S., et al. (2011). RUNX1 regulates corepressor interactions of PU.1. *Blood*, 117(24), 6498-6508. doi: 10.1182/blood-2010-10-312512
- Huang da, W., Sherman, B. T. and Lempicki, R. A. (2009a). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res*, 37(1), 1-13. doi: 10.1093/nar/gkn923
- Huang da, W., Sherman, B. T. and Lempicki, R. A. (2009b). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc*, 4(1), 44-57. doi: 10.1038/nprot.2008.211
- Huang, F. W., Hodis, E., Xu, M. J., Kryukov, G. V., Chin, L. and Garraway, L. A. (2013). Highly recurrent TERT promoter mutations in human melanoma. *Science*, 339(6122), 957-959. doi: 10.1126/science.1229259
- Huang, J., Nueda, A., Yoo, S. and Dynan, W. S. (1997). Heat Shock Transcription Factor 1 Binds Selectively in Vitro to Ku Protein and the Catalytic Subunit of the DNA-dependent Protein Kinase. *Journal of Biological Chemistry*, 272(41), 26009-26016. doi: 10.1074/jbc.272.41.26009
- Hulsen, T., de Vlieg, J. and Alkema, W. (2008). BioVenn – a web application for the comparison and visualization of biological lists using area-proportional Venn diagrams. *BMC Genomics*, 9(1), 488. doi: 10.1186/1471-2164-9-488
- Hurtado, A., Holmes, K. A., Ross-Innes, C. S., Schmidt, D. and Carroll, J. S. (2011). FOXA1 is a key determinant of estrogen receptor function and endocrine response. *Nat Genet*, 43(1), 27-33. doi: 10.1038/ng.730
- Iijima, S., Teraoka, H., Date, T. and Tsukada, K. (1992). DNA-activated protein kinase in Raji Burkitt's lymphoma cells. *Eur J Biochem*, 206(2), 595-603. doi: 10.1111/j.1432-1033.1992.tb16964.x
- Jorns, E., Lord, C. J. and Ashworth, A. (2009). Parallel RNAi and compound screens identify the PDK1 pathway as a

- target for tamoxifen sensitization. *Biochem J*, 417(1), 361-370. doi: 10.1042/BJ20081682
- Ito, S., D'Alessio, A. C., Taranova, O. V., Hong, K., Sowers, L. C. and Zhang, Y. (2010). Role of Tet proteins in 5mC to 5hmC conversion, ES-cell self-renewal and inner cell mass specification. *Nature*, 466(7310), 1129-1133. doi: 10.1038/nature09303
- Iwafuchi-Doi, M., Donahue, G., Kakumanu, A., Watts, J. A., Mahony, S., Pugh, B. F., *et al.* (2016). The Pioneer Transcription Factor FoxA Maintains an Accessible Nucleosome Configuration at Enhancers for Tissue-Specific Gene Activation. *Mol Cell*, 62(1), 79-91. doi: 10.1016/j.molcel.2016.03.001
- Iwafuchi-Doi, M. and Zaret, K. S. (2014). Pioneer transcription factors in cell reprogramming. *Genes Dev*, 28(24), 2679-2692. doi: 10.1101/gad.253443.114
- Izhar, L., Adamson, B., Ciccio, A., Lewis, J., Pontano-Vaites, L., Leng, Y., *et al.* (2015). A Systematic Analysis of Factors Localized to Damaged Chromatin Reveals PARP-Dependent Recruitment of Transcription Factors. *Cell Reports*, 11(9), 1486-1500. doi: 10.1016/j.celrep.2015.04.053
- Jackson, S. P., MacDonald, J. J., Lees-Miller, S. and Tjian, R. (1990). GC box binding induces phosphorylation of Sp1 by a DNA-dependent protein kinase. *Cell*, 63(1), 155-165.
- Jacobsen, B. M. and Horwitz, K. B. (2012). Progesterone receptors, their isoforms and progesterone regulated transcription. *Mol Cell Endocrinol*, 357(1-2), 18-29. doi: 10.1016/j.mce.2011.09.016
- Jeggo, P. A., Pearl, L. H. and Carr, A. M. (2016). DNA repair, genome stability and cancer: a historical perspective. *Nat Rev Cancer*, 16(1), 35-42. doi: 10.1038/nrc.2015.4
- Jette, N. and Lees-Miller, S. P. (2015). The DNA-dependent protein kinase: A multifunctional protein kinase with roles in DNA double strand break repair and mitosis. *Prog Biophys Mol Biol*, 117(2-3), 194-205. doi: 10.1016/j.pbiomolbio.2014.12.003
- Jeyakumar, M., Liu, X. F., Erdjument-Bromage, H., Tempst, P. and Bagchi, M. K. (2007). Phosphorylation of thyroid hormone receptor-associated nuclear receptor corepressor holocomplex by the DNA-dependent protein kinase enhances its histone deacetylase activity. *J Biol Chem*, 282(13), 9312-9322. doi: 10.1074/jbc.M609009200
- Jin, B., Jin, H. and Wang, J. (2017). Silencing of Interferon-Induced Transmembrane Protein 1 (IFITM1) Inhibits Proliferation, Migration, and Invasion in Lung Cancer Cells. *Oncol Res*. doi: 10.3727/096504017X14844360974116
- Jin, C., Zang, C., Wei, G., Cui, K., Peng, W., Zhao, K., *et al.* (2009). H3.3/H2A.Z double variant-containing nucleosomes mark 'nucleosome-free regions' of active promoters and other regulatory regions. *Nat Genet*, 41(8), 941-945. doi: 10.1038/ng.409
- John, S., Sabo, P. J., Thurman, R. E., Sung, M. H., Biddie, S. C., Johnson, T. A., *et al.* (2011). Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. *Nat Genet*, 43(3), 264-268. doi: 10.1038/ng.759
- Johnston, S. J. and Carroll, J. S. (2015). Transcription factors and chromatin proteins as therapeutic targets in cancer. *Biochim Biophys Acta*, 1855(2), 183-192. doi: 10.1016/j.bbcan.2015.02.002
- Johnston, S. R. D., Head, J., Pancholi, S., Detre, S., Martin, L.-A., Smith, I. E., *et al.* (2003). Integration of Signal Transduction Inhibitors with Endocrine Therapy. *An Approach to Overcoming Hormone Resistance in Breast Cancer*, 9(1), 524s-532s.
- Johnston, S. R. D., Kilburn, L. S., Ellis, P., Dodwell, D., Cameron, D., Hayward, L., *et al.* Fulvestrant plus anastrozole or placebo versus exemestane alone after progression on non-steroidal aromatase inhibitors in postmenopausal patients with hormone-receptor-positive locally advanced or metastatic breast cancer (SoFEA): a composite, multicentre, phase 3 randomised trial. *The Lancet Oncology*, 14(10), 989-998. doi: 10.1016/S1470-2045(13)70322-X
- Jolma, A., Yin, Y., Nitta, K. R., Dave, K., Popov, A., Taipale, M., *et al.* (2015). DNA-dependent formation of transcription factor pairs alters their binding specificity. *Nature*, 527(7578), 384-388. doi: 10.1038/nature15518
- Jones, P. A. (2012). Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat Rev Genet*, 13(7), 484-492. doi: 10.1038/nrg3230
- Jorruiz, S. M. and Bourdon, J. C. (2016). p53 Isoforms: Key Regulators of the Cell Fate Decision. *Cold Spring Harb Perspect Med*, 6(8). doi: 10.1101/cshperspect.a026039
- Jozwik, K. M., Chernukhin, I., Serandour, A. A., Nagarajan, S. and Carroll, J. S. (2016). FOXA1 Directs H3K4 Monomethylation at Enhancers via Recruitment of the Methyltransferase MLL3. *Cell Rep*, 17(10), 2715-2723. doi: 10.1016/j.celrep.2016.11.028
- Ju, B. G., Lunyak, V. V., Perissi, V., Garcia-Bassets, I., Rose, D. W., Glass, C. K., *et al.* (2006). A topoisomerase IIbeta-mediated dsDNA break required for regulated transcription. *Science*, 312(5781), 1798-1802. doi: 10.1126/science.1127196
- Ju, B. G. and Rosenfeld, M. G. (2006). A breaking strategy for topoisomerase IIbeta/PARP-1-dependent regulated transcription. *Cell Cycle*, 5(22), 2557-2560. doi: 10.4161/cc.5.22.3497
- Kagey, M. H., Newman, J. J., Bilodeau, S., Zhan, Y., Orlando, D. A., van Berkum, N. L., *et al.* (2010). Mediator and cohesin connect gene expression and chromatin architecture. *Nature*, 467(7314), 430-435. doi: 10.1038/nature09380
- Kalyuga, M., Gallego-Ortega, D., Lee, H. J., Roden, D. L., Cowley, M. J., Caldon, C. E., *et al.* (2012). ELF5 Suppresses Estrogen Sensitivity and Underpins the Acquisition of Antiestrogen Resistance in Luminal Breast Cancer. *PLoS Biol*, 10(12), e1001461. doi: 10.1371/journal.pbio.1001461
- Kan, Z., Jaiswal, B. S., Stinson, J., Janakiraman, V., Bhatt, D., Stern, H. M., *et al.* (2010). Diverse somatic mutation patterns and pathway alterations in human cancers. *Nature*, 466(7308), 869-873. doi: 10.1038/nature09208
- Kang, G. Y., Pyun, B. J., Seo, H. R., Jin, Y. B., Lee, H. J., Lee, Y. J., *et al.* (2013). Inhibition of Snail1-DNA-PKcs protein-protein interface sensitizes cancer cells and inhibits tumor metastasis. *J Biol Chem*, 288(45), 32506-32516. doi: 10.1074/jbc.M113.479840
- Kar, A. and Gutierrez-Hartmann, A. (2013). Molecular mechanisms of ETS transcription factor-mediated tumorigenesis. *Crit Rev Biochem Mol Biol*, 48(6), 522-543. doi: 10.3109/10409238.2013.838202
- Karpova, A. Y., Trost, M., Murray, J. M., Cantley, L. C. and Howley, P. M. (2002). Interferon regulatory factor-3 is an in vivo target of DNA-PK. *Proc Natl Acad Sci U S A*, 99(5), 2818-2823. doi: 10.1073/pnas.052713899
- Kelemen, O., Convertini, P., Zhang, Z., Wen, Y., Shen, M., Falaleeva, M., *et al.* (2013). Function of alternative splicing. *Gene*, 514(1), 1-30. doi: 10.1016/j.gene.2012.07.083
- Keller, P. J., Lin, A. F., Arendt, L. M., Klebba, I., Jones, A. D., Rudnick, J. A., *et al.* (2010). Mapping the cellular and molecular heterogeneity of normal and malignant breast tissues and cultured cell lines. *Breast Cancer*

- Research : BCR, 12(5), R87-R87. doi: 10.1186/bcr2755
- Kendrick, H., Regan, J. L., Magnay, F.-A., Grigoriadis, A., Mitsopoulos, C., Zvelebil, M., *et al.* (2008). Transcriptome analysis of mammary epithelial subpopulations identifies novel determinants of lineage commitment and cell fate. *BMC Genomics*, 9(1), 591. doi: 10.1186/1471-2164-9-591
- Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., *et al.* (2002). The human genome browser at UCSC. *Genome Res*, 12(6), 996-1006. doi: 10.1101/gr.229102. Article published online before print in May 2002
- Khabar, K. S. (2017). Hallmarks of cancer and AU-rich elements. *Wiley Interdiscip Rev RNA*, 8(1). doi: 10.1002/wrna.1368
- Khanna, A. (2015). DNA damage in cancer therapeutics: a boon or a curse? *Cancer Res*, 75(11), 2133-2138. doi: 10.1158/0008-5472.CAN-14-3247
- Kim, N. H., Sung, H. Y., Choi, E. N., Lyu, D., Choi, H. J., Ju, W., *et al.* (2014). Aberrant DNA methylation in the IFITM1 promoter enhances the metastatic phenotype in an intraperitoneal xenograft model of human ovarian cancer. *Oncol Rep*, 31(5), 2139-2146. doi: 10.3892/or.2014.3110
- Kim, S., Brostromer, E., Xing, D., Jin, J., Chong, S., Ge, H., *et al.* (2013). Probing allostery through DNA. *Science*, 339(6121), 816-819. doi: 10.1126/science.1229223
- Kinnaird, A., Zhao, S., Wellen, K. E. and Michelakis, E. D. (2016). Metabolic control of epigenetics in cancer. *Nat Rev Cancer*, 16(11), 694-707. doi: 10.1038/nrc.2016.82
- Kinsella, R. J., Kahari, A., Haider, S., Zamora, J., Proctor, G., Spudich, G., *et al.* (2011). Ensembl BioMart: a hub for data retrieval across taxonomic space. *Database (Oxford)*, 2011, bar030. doi: 10.1093/database/bar030
- Koch, H., Zhang, R., Verdoodt, B., Bailey, A., Zhang, C.-D., Yates, J. R., *et al.* (2007). Large-Scale Identification of c-MYC-Associated Proteins Using a Combined TAP/MudPIT Approach. *Cell Cycle*, 6(2), 205-217. doi: 10.4161/cc.6.2.3742
- Kooistra, S. M. and Helin, K. (2012). Molecular mechanisms and potential functions of histone demethylases. *Nat Rev Mol Cell Biol*, 13(5), 297-311. doi: 10.1038/nrm3327
- Kornblihtt, A. R., Schor, I. E., Allo, M., Dujardin, G., Petrillo, E. and Munoz, M. J. (2013). Alternative splicing: a pivotal step between eukaryotic transcription and translation. *Nat Rev Mol Cell Biol*, 14(3), 153-165.
- Koster, M. J., Snel, B. and Timmers, H. T. (2015). Genesis of chromatin and transcription dynamics in the origin of species. *Cell*, 161(4), 724-736. doi: 10.1016/j.cell.2015.04.033
- Krogan, N. J., Kim, M., Tong, A., Golshani, A., Cagney, G., Canadien, V., *et al.* (2003). Methylation of histone H3 by Set2 in *Saccharomyces cerevisiae* is linked to transcriptional elongation by RNA polymerase II. *Mol Cell Biol*, 23(12), 4207-4218.
- Kruse, J.-P. and Gu, W. (2008). SnapShot: p53 Posttranslational Modifications. *Cell*, 133(5), 930-930.e931. doi: 10.1016/j.cell.2008.05.020
- Kuleshov, M. V., Jones, M. R., Rouillard, A. D., Fernandez, N. F., Duan, Q., Wang, Z., *et al.* (2016). Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res*, 44(W1), W90-97. doi: 10.1093/nar/gkw377
- Kurimasa, A., Kumano, S., Boubnov, N. V., Story, M. D., Tung, C. S., Peterson, S. R., *et al.* (1999). Requirement for the kinase activity of human DNA-dependent protein kinase catalytic subunit in DNA strand break rejoining. *Mol Cell Biol*, 19(5), 3877-3884.
- Kwok, H. F., Zhang, S. D., McCrudden, C. M., Yuen, H. F., Ting, K. P., Wen, Q., *et al.* (2015). Prognostic significance of minichromosome maintenance proteins in breast cancer. *Am J Cancer Res*, 5(1), 52-71.
- Laitem, C., Leprivier, G., Choul-Li, S., Begue, A., Monte, D., Larsimont, D., *et al.* (2009). Ets-1 p27: a novel Ets-1 isoform with dominant-negative effects on the transcriptional properties and the subcellular localization of Ets-1 p51. *Oncogene*, 28(20), 2087-2099. doi: 10.1038/nc.2009.72
- Lamb, R., Fiorillo, M., Chadwick, A., Ozsvari, B., Reeves, K. J., Smith, D. L., *et al.* (2015). Doxycycline down-regulates DNA-PK and radiosensitizes tumor initiating cells: Implications for more effective radiation therapy. *Oncotarget*, 6(16), 14005-14025. doi: 10.18632/oncotarget.4159
- Lamhamedi-Cherradi, S.-E., Menegaz, B. A., Ramamoorthy, V., Aiyer, R. A., Maywald, R. L., Buford, A. S., *et al.* (2015). An Oral Formulation of YK-4-279: Preclinical Efficacy and Acquired Resistance Patterns in Ewing Sarcoma. *Mol Cancer Ther*, 14(7), 1591-1604. doi: 10.1158/1535-7163.mct-14-0334
- Lamonica, J. M., Deng, W., Kadauke, S., Campbell, A. E., Gamsjaeger, R., Wang, H., *et al.* (2011). Bromodomain protein Brd3 associates with acetylated GATA1 to promote its chromatin occupancy at erythroid target genes. *Proc Natl Acad Sci U S A*, 108(22), E159-168. doi: 10.1073/pnas.1102140108
- Langmead, B., Trapnell, C., Pop, M. and Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*, 10(3), R25. doi: 10.1186/gb-2009-10-3-r25
- Lapinskas, E. J., Palmer, J., Ricardo, S., Hertzog, P. J., Hammacher, A. and Pritchard, M. A. (2004). A major site of expression of the ets transcription factor Elf5 is epithelia of exocrine glands. *Histochem Cell Biol*, 122(6), 521-526. doi: 10.1007/s00418-004-0713-x
- Lapinskas, E. J., Svobodova, S., Davis, I. D., Cebon, J., Hertzog, P. J. and Pritchard, M. A. (2011). The Ets transcription factor ELF5 functions as a tumor suppressor in the kidney. *Twin Res Hum Genet*, 14(4), 316-322. doi: 10.1375/twin.14.4.316
- Latchman, D. S. (2001). Transcription factors: bound to activate or repress. *Trends Biochem Sci*, 26(4), 211-213.
- Latos, P. A., Sienerth, A. R., Murray, A., Senner, C. E., Muto, M., Ikawa, M., *et al.* (2015). Elf5-centered transcription factor hub controls trophoblast stem cell self-renewal and differentiation through stoichiometry-sensitive shifts in target gene networks. *Genes Dev*, 29(23), 2435-2448. doi: 10.1101/gad.268821.115
- Lavin, M. F. (2007). ATM and the Mre11 complex combine to recognize and signal DNA double-strand breaks. *Oncogene*, 26(56), 7749-7758.
- Law, A. M., Lim, E., Ormandy, C. J. and Gallego-Ortega, D. (2017a). The innate and adaptive infiltrating immune systems as targets for breast cancer immunotherapy. *Endocr Relat Cancer*, 24(4), R123-R144. doi: 10.1530/ERC-16-0404
- Law, A. M. K., Yin, J. X. M., Castillo, L., Young, A. I. J., Piggan, C., Rogers, S., *et al.* (2017b). Andy's Algorithms: new automated digital image analysis pipelines for FIJI. *Sci Rep*, 7(1), 15717. doi: 10.1038/s41598-017-15885-6
- Law, C. W., Chen, Y., Shi, W. and Smyth, G. K. (2014). voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol*, 15(2), R29. doi: 10.1186/gb-2014-15-2-r29
- Lawrence, M. S., Stojanov, P., Mermel, C. H., Robinson, J. T., Garraway, L. A., Golub, T. R., *et al.* (2014). Discovery

- and saturation analysis of cancer genes across 21 tumour types. *Nature*, 505(7484), 495-501. doi: 10.1038/nature12912
- Lazo, J. S. and Sharlow, E. R. (2016). Drugging Undruggable Molecular Cancer Targets. *Annu Rev Pharmacol Toxicol*, 56, 23-40. doi: 10.1146/annurev-pharmtox-010715-103440
- Lee, D. Y., Hayes, J. J., Pruss, D. and Wolffe, A. P. (1993). A positive role for histone acetylation in transcription factor access to nucleosomal DNA. *Cell*, 72(1), 73-84.
- Lee, G. M., Donaldson, L. W., Pufall, M. A., Kang, H. S., Pot, I., Graves, B. J., et al. (2005a). The structural and dynamic basis of Ets-1 DNA binding autoinhibition. *J Biol Chem*, 280(8), 7088-7099. doi: 10.1074/jbc.M410722200
- Lee, H. J., Gallego-Ortega, D., Ledger, A., Schramek, D., Joshi, P., Szwarc, M. M., et al. (2013). Progesterone drives mammary secretory differentiation via RankL-mediated induction of Elf5 in luminal progenitor cells. *Development*, 140(7), 1397-1401. doi: 10.1242/dev.088948
- Lee, H. J., Hinshelwood, R. A., Bouras, T., Gallego-Ortega, D., Valdes-Mora, F., Blazek, K., et al. (2011a). Lineage specific methylation of the Elf5 promoter in mammary epithelial cells. *Stem Cells*, 29(10), 1611-1619. doi: 10.1002/stem.706
- Lee, H. S., Choe, G., Park, K. U., Park, D. J., Yang, H. K., Lee, B. L., et al. (2007). Altered expression of DNA-dependent protein kinase catalytic subunit (DNA-PKcs) during gastric carcinogenesis and its clinical implications on gastric cancer. *Int J Oncol*, 31(4), 859-866.
- Lee, H. S., Yang, H. K., Kim, W. H. and Choe, G. (2005b). Loss of DNA-dependent protein kinase catalytic subunit (DNA-PKcs) expression in gastric cancers. *Cancer Res Treat*, 37(2), 98-102. doi: 10.4143/crt.2005.37.2.98
- Lee, J., Goh, S. H., Song, N., Hwang, J. A., Nam, S., Choi, I. J., et al. (2012). Overexpression of IFITM1 has clinicopathologic effects on gastric cancer and is regulated by an epigenetic mechanism. *Am J Pathol*, 181(1), 43-52. doi: 10.1016/j.ajpath.2012.03.027
- Lee, K. J., Lin, Y. F., Chou, H. Y., Yajima, H., Fattah, K. R., Lee, S. C., et al. (2011b). Involvement of DNA-dependent protein kinase in normal cell cycle progression through mitosis. *J Biol Chem*, 286(14), 12796-12802. doi: 10.1074/jbc.M110.212969
- Lee, S.-w., Cho, K.-J., Park, J.-h., Kim, S. Y., Nam, S. Y., Lee, B.-J., et al. (2005c). Expressions of Ku70 and DNA-PKcs as prognostic indicators of local control in nasopharyngeal carcinoma. *International Journal of Radiation Oncology\*Biophysics*, 62(5), 1451-1457. doi: 10.1016/j.ijrobp.2004.12.049
- Lee, T. I., Rinaldi, N. J., Robert, F., Odom, D. T., Bar-Joseph, Z., Gerber, G. K., et al. (2002). Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science*, 298(5594), 799-804. doi: 10.1126/science.1075090
- Lee, T. I. and Young, R. A. (2013). Transcriptional regulation and its misregulation in disease. *Cell*, 152(6), 1237-1251. doi: 10.1016/j.cell.2013.02.014
- Lefstin, J. A. and Yamamoto, K. R. (1998). Allosteric effects of DNA on transcriptional regulators. *Nature*, 392(6679), 885-888.
- Lehmann, B. D., Bauer, J. A., Chen, X., Sanders, M. E., Chakravarthy, A. B., Shyr, Y., et al. (2011). Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. *J Clin Invest*, 121(7), 2750-2767. doi: 10.1172/JCI45014
- Lelli, K. M., Slattery, M. and Mann, R. S. (2012). Disentangling the many layers of eukaryotic transcriptional regulation. *Annu Rev Genet*, 46, 43-68. doi: 10.1146/annurev-genet-110711-155437
- Leung, T. H., Hoffmann, A. and Baltimore, D. (2004). One Nucleotide in a KB Site Can Determine Cofactor Specificity for NF-KB Dimers. *Cell*, 118(4), 453-464. doi: 10.1016/j.cell.2004.08.007
- Levin, E. R. and Hammes, S. R. (2016). Nuclear receptors outside the nucleus: extranuclear signalling by steroid receptors. *Nat Rev Mol Cell Biol*, 17(12), 783-797. doi: 10.1038/nrm.2016.122
- Levine, M. (2010). Transcriptional enhancers in animal development and evolution. *Curr Biol*, 20(17), R754-763. doi: 10.1016/j.cub.2010.06.070
- Levine, M. and Tjian, R. (2003). Transcription regulation and animal diversity. *Nature*, 424(6945), 147-151. doi: 10.1038/nature01763
- Lewis-Wambi, J. S., Cunliffe, H. E., Kim, H. R., Willis, A. L. and Jordan, V. C. (2008). Overexpression of CEACAM6 promotes migration and invasion of oestrogen-deprived breast cancer cells. *Eur J Cancer*, 44(12), 1770-1779. doi: 10.1016/j.ejca.2008.05.016
- Li, B., Carey, M. and Workman, J. L. (2007a). The role of chromatin during transcription. *Cell*, 128(4), 707-719. doi: 10.1016/j.cell.2007.01.015
- Li, B. and Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, 12, 323. doi: 10.1186/1471-2105-12-323
- Li, H., Fischle, W., Wang, W., Duncan, E. M., Liang, L., Murakami-Ishibe, S., et al. (2007b). Structural basis for lower lysine methylation state-specific readout by MBT repeats of L3MBTL1 and an engineered PHD finger. *Mol Cell*, 28(4), 677-691. doi: 10.1016/j.molcel.2007.10.023
- Li, K., Guo, Y., Yang, X., Zhang, Z., Zhang, C. and Xu, Y. (2017). ELF5-Mediated AR Activation Regulates Prostate Cancer Progression. *Sci Rep*, 7, 42759. doi: 10.1038/srep42759
- Li, R., Pei, H. and Watson, D. K. (2000). Regulation of Ets function by protein - protein interactions. *Oncogene*, 19(55), 6514-6523. doi: 10.1038/sj.onc.1204035
- Li, W., Notani, D., Ma, Q., Tanasa, B., Nunez, E., Chen, A. Y., et al. (2013). Functional roles of enhancer RNAs for oestrogen-dependent transcriptional activation. *Nature*, 498(7455), 516-520. doi: 10.1038/nature12210
- Li, Z., Owonikoko, T. K., Sun, S. Y., Ramalingam, S. S., Doetsch, P. W., Xiao, Z. Q., et al. (2012). c-Myc suppression of DNA double-strand break repair. *Neoplasia*, 14(12), 1190-1202.
- Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J. P. and Tamayo, P. (2015). The Molecular Signatures Database Hallmark Gene Set Collection. *Cell Systems*, 1(6), 417-425. doi: 10.1016/j.cels.2015.12.004
- Liiv, I., Rebane, A., Org, T., Saare, M., Maslovskaja, J., Kisand, K., et al. (2008). DNA-PK contributes to the phosphorylation of AIRE: Importance in transcriptional activity. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research*, 1783(1), 74-83. doi: 10.1016/j.bbamcr.2007.09.003
- Lim, E., Vaillant, F., Wu, D., Forrest, N. C., Pal, B., Hart, A. H., et al. (2009). Aberrant luminal progenitors as the candidate target population for basal tumor development in BRCA1 mutation carriers. *Nat Med*, 15(8), 907-913. doi: 10.1038/nm.2000
- Lin, C. Y., Loven, J., Rahl, P. B., Paranal, R. M., Burge, C. B., Bradner, J. E., et al. (2012). Transcriptional amplification in tumor cells with elevated c-Myc. *Cell*, 151(1), 56-67. doi: 10.1016/j.cell.2012.08.026

- Lin, Y., Wu, Y., Li, J., Dong, C., Ye, X., Chi, Y. I., *et al.* (2010). The SNAG domain of Snail1 functions as a molecular hook for recruiting lysine-specific demethylase 1. *EMBO J*, 29(11), 1803-1816. doi: 10.1038/emboj.2010.63
- Lindeman, G. J., Wittlin, S., Lada, H., Naylor, M. J., Santamaria, M., Zhang, J. G., *et al.* (2001). SOCS1 deficiency results in accelerated mammary gland development and rescues lactation in prolactin receptor-deficient mice. *Genes Dev*, 15(13), 1631-1636. doi: 10.1101/gad.880801
- Liu, F., Wang, L., Perna, F. and Nimer, S. D. (2016). Beyond transcription factors: how oncogenic signalling reshapes the epigenetic landscape. *Nat Rev Cancer*, 16(6), 359-372. doi: 10.1038/nrc.2016.41
- Liu, S., Opiyo, S. O., Manthey, K., Glanzer, J. G., Ashley, A. K., Amerin, C., *et al.* (2012). Distinct roles for DNA-PK, ATM and ATR in RPA phosphorylation and checkpoint activation in response to replication stress. *Nucleic Acids Research*, 40(21), 10780-10794. doi: 10.1093/nar/gks849
- Liu, W. L., Coleman, R. A., Ma, E., Grob, P., Yang, J. L., Zhang, Y., *et al.* (2009). Structures of three distinct activator-TFIIID complexes. *Genes Dev*, 23(13), 1510-1521. doi: 10.1101/gad.1790709
- Liu, X. S., Chandramouly, G., Rass, E., Guan, Y., Wang, G., Hobbs, R. M., *et al.* (2015). LRF maintains genome integrity by regulating the non-homologous end joining pathway of DNA repair. *Nat Commun*, 6, 8325. doi: 10.1038/ncomms9325
- Liu, Z., Merkurjev, D., Yang, F., Li, W., Oh, S., Friedman, M. J., *et al.* (2014). Enhancer activation requires trans-recruitment of a mega transcription factor complex. *Cell*, 159(2), 358-373. doi: 10.1016/j.cell.2014.08.027
- Lock, L. F., Takagi, N. and Martin, G. R. (1987). Methylation of the Hprt gene on the inactive X occurs after chromosome inactivation. *Cell*, 48(1), 39-46. doi: 10.1016/0092-8674(87)90353-9
- Long, H. K., Blackledge, N. P. and Klose, R. J. (2013). ZF-CxxC domain-containing proteins, CpG islands and the chromatin connection. *Biochem Soc Trans*, 41(3), 727-740. doi: 10.1042/BST20130028
- Lord, C. J. and Ashworth, A. (2012). The DNA damage response and cancer therapy. *Nature*, 481(7381), 287-294. doi: 10.1038/nature10760
- Loven, J., Hoke, H. A., Lin, C. Y., Lau, A., Orlando, D. A., Vakoc, C. R., *et al.* (2013). Selective inhibition of tumor oncogenes by disruption of super-enhancers. *Cell*, 153(2), 320-334. doi: 10.1016/j.cell.2013.03.036
- Lui, A. J., Geanes, E. S., Ogony, J., Behbod, F., Marquess, J., Valdez, K., *et al.* (2017). IFITM1 suppression blocks proliferation and invasion of aromatase inhibitor-resistant breast cancer in vivo by JAK/STAT-mediated induction of p21. *Cancer Lett*, 399, 29-43. doi: 10.1016/j.canlet.2017.04.005
- Lupien, M., Eeckhoutte, J., Meyer, C. A., Wang, Q., Zhang, Y., Li, W., *et al.* (2008). FoxA1 translates epigenetic signatures into enhancer-driven lineage-specific transcription. *Cell*, 132(6), 958-970. doi: 10.1016/j.cell.2008.01.018
- Lupien, M., Meyer, C. A., Bailey, S. T., Eeckhoutte, J., Cook, J., Westerling, T., *et al.* (2010). Growth factor stimulation induces a distinct ER(alpha) cisome underlying breast cancer endocrine resistance. *Genes Dev*, 24(19), 2219-2227. doi: 10.1101/gad.1944810
- Machanick, P. and Bailey, T. L. (2011). MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics*, 27(12), 1696-1697. doi: 10.1093/bioinformatics/btr189
- Mackereth, C. D., Scharpf, M., Gentile, L. N., MacIntosh, S. E., Slupsky, C. M. and McIntosh, L. P. (2004). Diversity in structure and function of the Ets family PNT domains. *J Mol Biol*, 342(4), 1249-1264. doi: 10.1016/j.jmb.2004.07.094
- Maggi, A. (2011). Liganded and unliganded activation of estrogen receptor and hormone replacement therapies. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*, 1812(8), 1054-1060. doi: 10.1016/j.bbdis.2011.05.001
- Mahaney, B. L., Meek, K. and Lees-Miller, S. P. (2009). Repair of ionizing radiation-induced DNA double-strand breaks by non-homologous end-joining. *Biochem J*, 417(3), 639-650. doi: 10.1042/BJ20080413
- Malewicz, M., Kadkhodaei, B., Kee, N., Volakakis, N., Hellman, U., Viktorsson, K., *et al.* (2011). Essential role for DNA-PK-mediated phosphorylation of NR4A nuclear orphan receptors in DNA double-strand break repair. *Genes Dev*, 25(19), 2031-2040. doi: 10.1101/gad.16872411
- Manavathi, B., Rayala, S. K. and Kumar, R. (2007). Phosphorylation-dependent regulation of stability and transforming potential of ETS transcriptional factor ESE-1 by p21-activated kinase 1. *J Biol Chem*, 282(27), 19820-19830. doi: 10.1074/jbc.M702309200
- Manavathi, B., Samanthapudi, V. S. and Gajulapalli, V. N. (2014). Estrogen receptor coregulators and pioneer factors: the orchestrators of mammary gland cell fate and development. *Front Cell Dev Biol*, 2, 34. doi: 10.3389/fcell.2014.00034
- Manning, K. S. and Cooper, T. A. (2017). The roles of RNA processing in translating genotype to phenotype. *Nat Rev Mol Cell Biol*, 18(2), 102-114. doi: 10.1038/nrm.2016.139
- Mantione, K. J., Cream, R. M., Kuzelova, H., Ptacek, R., Raboch, J., Samuel, J. M., *et al.* (2014). Comparing bioinformatic gene expression profiling methods: microarray and RNA-Seq. *Med Sci Monit Basic Res*, 20, 138-142. doi: 10.12659/MSMBR.892101
- MAQC Consortium. (2006). The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat Biotech*, 24(9), 1151-1161. doi: 10.1038/nbt1239
- Maraqa, L., Cummings, M., Peter, M. B., Shaaban, A. M., Horgan, K., Hanby, A. M., *et al.* (2008). Carcinoembryonic antigen cell adhesion molecule 6 predicts breast cancer recurrence following adjuvant tamoxifen. *Clin Cancer Res*, 14(2), 405-411. doi: 10.1158/1078-0432.CCR-07-1363
- Marmorstein, R. and Zhou, M. M. (2014). Writers and readers of histone acetylation: structure, mechanism, and inhibition. *Cold Spring Harb Perspect Biol*, 6(7), a018762. doi: 10.1101/cshperspect.a018762
- Marmot, M. G., Altman, D. G., Cameron, D. A., Dewar, J. A., Thompson, S. G. and Wilcox, M. (2013). The benefits and harms of breast cancer screening: an independent review. *Br J Cancer*, 108, 2205. doi: 10.1038/bjc.2013.177
- Marshman, E., Green, K. A., Flint, D. J., White, A., Streuli, C. H. and Westwood, M. (2003). Insulin-like growth factor binding protein 5 and apoptosis in mammary epithelial cells. *J Cell Sci*, 116(4), 675-682. doi: 10.1242/jcs.00263
- Martin, L. A., Ghazoui, Z., Weigel, M. T., Pancholi, S., Dunbier, A., Johnston, S., *et al.* (2011). An in vitro model showing adaptation to long-term oestrogen deprivation highlights the clinical potential for targeting kinase pathways in combination with aromatase inhibition. *Steroids*, 76(8), 772-776. doi: 10.1016/j.steroids.2011.02.035
- Mathelier, A., Fornes, O., Arenillas, D. J., Chen, C. Y., Denay, G., Lee, J., *et al.* (2016). JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res*, 44(D1), D110-115. doi: 10.1093/nar/gkv1176

- Mathieu, A. L., Verronese, E., Rice, G. I., Fouyssac, F., Bertrand, Y., Picard, C., *et al.* (2015). PRKDC mutations associated with immunodeficiency, granuloma, and autoimmune regulator-dependent autoimmunity. *J Allergy Clin Immunol*, 135(6), 1578-1588 e1575. doi: 10.1016/j.jaci.2015.01.040
- Maurano, M. T., Wang, H., John, S., Shafer, A., Canfield, T., Lee, K., *et al.* (2015). Role of DNA Methylation in Modulating Transcription Factor Occupancy. *Cell Rep*, 12(7), 1184-1195. doi: 10.1016/j.celrep.2015.07.024
- Mayeur, G. L., Kung, W. J., Martinez, A., Izumiya, C., Chen, D. J. and Kung, H. J. (2005). Ku is a novel transcriptional recycling coactivator of the androgen receptor in prostate cancer cells. *J Biol Chem*, 280(11), 10827-10833. doi: 10.1074/jbc.M413336200
- Maze, I., Noh, K. M., Soshnev, A. A. and Allis, C. D. (2014). Every amino acid matters: essential contributions of histone variants to mammalian development and disease. *Nat Rev Genet*, 15(4), 259-271. doi: 10.1038/nrg3673
- McCarthy, D. J., Chen, Y. and Smyth, G. K. (2012). Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res*, 40(10), 4288-4297. doi: 10.1093/nar/gks042
- McLean, C. Y., Bristor, D., Hiller, M., Clarke, S. L., Schaar, B. T., Lowe, C. B., *et al.* (2010). GREAT improves functional interpretation of cis-regulatory regions. *Nat Biotechnol*, 28(5), 495-501. doi: 10.1038/nbt.1630
- McNeil, C. M., Sergio, C. M., Anderson, L. R., Inman, C. K., Eggleton, S. A., Murphy, N. C., *et al.* (2006). c-Myc overexpression and endocrine resistance in breast cancer. *The Journal of Steroid Biochemistry and Molecular Biology*, 102(1), 147-155. doi: 10.1016/j.jsmb.2006.09.028
- Medunjanin, S., Weinert, S., Poitz, D., Schmeisser, A., Strasser, R. H. and Braun-Dullaeus, R. C. (2010a). Transcriptional activation of DNA-dependent protein kinase catalytic subunit gene expression by oestrogen receptor- $\alpha$ . *EMBO Rep*, 11(3), 208-213. doi: 10.1038/embor.2009.279
- Medunjanin, S., Weinert, S., Schmeisser, A., Mayer, D. and Braun-Dullaeus, R. C. (2010b). Interaction of the double-strand break repair kinase DNA-PK and estrogen receptor- $\alpha$ . *Mol Biol Cell*, 21(9), 1620-1628. doi: 10.1091/mbc.E09-08-0724
- Meek, D. W. and Anderson, C. W. (2009). Posttranslational modification of p53: cooperative integrators of function. *Cold Spring Harb Perspect Biol*, 1(6), a000950. doi: 10.1101/cshperspect.a000950
- Meek, K., Lees-Miller, S. P. and Modesti, M. (2012). N-terminal constraint activates the catalytic subunit of the DNA-dependent protein kinase in the absence of DNA or Ku. *Nucleic Acids Research*, 40(7), 2964-2973. doi: 10.1093/nar/gkr1211
- Meier, K. and Brehm, A. (2014). Chromatin regulation: how complex does it get? *Epigenetics*, 9(11), 1485-1495. doi: 10.4161/15592294.2014.971580
- Meijsing, S. H., Pufall, M. A., So, A. Y., Bates, D. L., Chen, L. and Yamamoto, K. R. (2009). DNA binding site sequence directs glucocorticoid receptor structure and activity. *Science*, 324(5925), 407-410. doi: 10.1126/science.1164265
- Melchor, L., Molyneux, G., Mackay, A., Magnay, F.-A., Atienza, M., Kendrick, H., *et al.* (2014). Identification of cellular and genetic drivers of breast cancer heterogeneity in genetically engineered mouse tumour models. *J Pathol*, 233(2), 124-137. doi: 10.1002/path.4345
- Mellacheruvu, D., Wright, Z., Couzens, A. L., Lambert, J.-P., St-Denis, N. A., Li, T., *et al.* (2013). The CRAPome: a contaminant repository for affinity purification-mass spectrometry data. *Nat Meth*, 10(8), 730-736. doi: 10.1038/nmeth.2557
- Merico, D., Isserlin, R., Stueker, O., Emili, A. and Bader, G. D. (2010). Enrichment Map: A Network-Based Method for Gene-Set Enrichment Visualization and Interpretation. *PLoS One*, 5(11), e13984. doi: 10.1371/journal.pone.0013984
- Mertins, P., Mani, D. R., Ruggles, K. V., Gillette, M. A., Clauser, K. R., Wang, P., *et al.* (2016). Proteogenomics connects somatic mutations to signalling in breast cancer. *Nature*, 534(7605), 55-62. doi: 10.1038/nature18003
- Mesquita, B., Lopes, P., Rodrigues, A., Pereira, D., Afonso, M., Leal, C., *et al.* (2013). Frequent copy number gains at 1q21 and 1q32 are associated with overexpression of the ETS transcription factors ETV3 and ELF3 in breast cancer irrespective of molecular subtypes. *Breast Cancer Res Treat*, 138(1), 37-45. doi: 10.1007/s10549-013-2408-2
- Metser, G., Shin, H. Y., Wang, C., Yoo, K. H., Oh, S., Villarino, A. V., *et al.* (2016). An autoregulatory enhancer controls mammary-specific STAT5 functions. *Nucleic Acids Res*, 44(3), 1052-1063. doi: 10.1093/nar/gkv999
- Metzger, D. E., Stahlman, M. T. and Shannon, J. M. (2008). Misexpression of ELF5 disrupts lung branching and inhibits epithelial differentiation. *Dev Biol*, 320(1), 149-160. doi: 10.1016/j.ydbio.2008.04.038
- Meyer, N. and Penn, L. Z. (2008). Reflecting on 25 years with MYC. *Nat Rev Cancer*, 8(12), 976-990.
- Meyer, T. and Vinkemeier, U. (2004). Nucleocytoplasmic shuttling of STAT transcription factors. *Eur J Biochem*, 271(23-24), 4606-4612. doi: 10.1111/j.1432-1033.2004.04423.x
- Millard, C. J., Watson, P. J., Fairall, L. and Schwabe, J. W. (2013). An evolving understanding of nuclear receptor coregulator proteins. *J Mol Endocrinol*, 51(3), T23-36. doi: 10.1530/JME-13-0227
- Miller, J. A. and Widom, J. (2003). Collaborative competition mechanism for gene activation in vivo. *Mol Cell Biol*, 23(5), 1623-1632.
- Miller, J. L. and Grant, P. A. (2013). The role of DNA methylation and histone modifications in transcriptional regulation in humans. In T. K. Kundu (Ed.), *Epigenetics: Development and Disease* (Vol. 61, pp. 289-317): Springer.
- Miller, K. E., Kim, Y., Huh, W.-K. and Park, H.-O. (2015). Bimolecular Fluorescence Complementation (BiFC) Analysis: Advances and Recent Applications for Genome-Wide Interaction Studies. *Journal of Molecular Biology*, 427(11), 2039-2055. doi: 10.1016/j.jmb.2015.03.005
- Miller, T. W., Balko, J. M., Ghazoui, Z., Dunbier, A., Anderson, H., Dowsett, M., *et al.* (2011). A gene expression signature from human breast cancer cells with acquired hormone independence identifies MYC as a mediator of antiestrogen resistance. *Clin Cancer Res*, 17(7), 2024-2034. doi: 10.1158/1078-0432.CCR-10-2567
- Millour, J., Constantinidou, D., Stavropoulou, A. V., Wilson, M. S., Myatt, S. S., Kwok, J. M., *et al.* (2010). FOXM1 is a transcriptional target of ER $\alpha$  and has a critical role in breast cancer endocrine sensitivity and resistance. *Oncogene*, 29(20), 2983-2995. doi: 10.1038/onc.2010.47
- Miranda, T. B., Voss, T. C., Sung, M. H., Baek, S., John, S., Hawkins, M., *et al.* (2013). Reprogramming the chromatin landscape: interplay of the estrogen and glucocorticoid receptors at the genomic level. *Cancer Res*, 73(16), 5130-5139. doi: 10.1158/0008-5472.CAN-13-0742
- Mirny, L. A. (2010). Nucleosome-mediated cooperativity between transcription factors. *Proc Natl Acad Sci U S A*, 107(52), 22534-22539. doi: 10.1073/pnas.0913805107



- Mohammed, H., D'Santos, C., Serandour, A. A., Ali, H. R., Brown, G. D., Atkins, A., *et al.* (2013). Endogenous purification reveals GREB1 as a key estrogen receptor regulatory factor. *Cell Rep*, 3(2), 342-349. doi: 10.1016/j.celrep.2013.01.010
- Mohammed, H., Russell, I. A., Stark, R., Rueda, O. M., Hickey, T. E., Tarulli, G. A., *et al.* (2015). Progesterone receptor modulates ERalpha action in breast cancer. *Nature*, 523(7560), 313-317. doi: 10.1038/nature14583
- Mohammed, H., Taylor, C., Brown, G. D., Papachristou, E. K., Carroll, J. S. and D'Santos, C. S. (2016). Rapid immunoprecipitation mass spectrometry of endogenous proteins (RIME) for analysis of chromatin complexes. *Nat Protoc*, 11(2), 316-326. doi: 10.1038/nprot.2016.020
- Molina, S., Guerif, S., Garcia, A., Debais, C., Irani, J. and Fromont, G. (2016). DNA-PKcs Expression Is a Predictor of Biochemical Recurrence After Permanent Iodine 125 Interstitial Brachytherapy for Prostate Cancer. *International Journal of Radiation Oncology\*Biophysics*, 95(3), 965-972. doi: 10.1016/j.ijrobp.2016.02.015
- Molyneux, G., Geyer, F. C., Magnay, F. A., McCarthy, A., Kendrick, H., Natrajan, R., *et al.* (2010). BRCA1 basal-like breast cancers originate from luminal epithelial progenitors and not from basal stem cells. *Cell Stem Cell*, 7(3), 403-417. doi: 10.1016/j.stem.2010.07.010
- Moriarty, K., Kim, K. H. and Bender, J. R. (2006). Minireview: estrogen receptor-mediated rapid signaling. *Endocrinology*, 147(12), 5557-5563. doi: 10.1210/en.2006-0729
- Morris, K. V. and Mattick, J. S. (2014). The rise of regulatory RNA. *Nat Rev Genet*, 15(6), 423-437. doi: 10.1038/nrg3722
- Mozzetta, C., Boyarchuk, E., Pontis, J. and Ait-Si-Ali, S. (2015). Sound of silence: the properties and functions of repressive Lys methyltransferases. *Nat Rev Mol Cell Biol*, 16(8), 499-513. doi: 10.1038/nrm4029
- Mueller-Planitz, F., Klinker, H. and Becker, P. B. (2013). Nucleosome sliding mechanisms: new twists in a looped history. *Nat Struct Mol Biol*, 20(9), 1026-1032. doi: 10.1038/nsmb.2648
- Musgrove, E. A., Sergio, C. M., Loi, S., Inman, C. K., Anderson, L. R., Alles, M. C., *et al.* (2008). Identification of Functional Networks of Estrogen- and c-Myc-Responsive Genes and Their Relationship to Response to Tamoxifen Therapy in Breast Cancer. *PLoS One*, 3(8), e2987. doi: 10.1371/journal.pone.0002987
- Musgrove, E. A. and Sutherland, R. L. (2009). Biological determinants of endocrine resistance in breast cancer. *Nat Rev Cancer*, 9(9), 631-643. doi: 10.1038/nrc2713
- Naftelberg, S., Schor, I. E., Ast, G. and Kornblitt, A. R. (2015). Regulation of alternative splicing through coupling with transcription and chromatin structure. *Annu Rev Biochem*, 84, 165-198. doi: 10.1146/annurev-biochem-060614-034242
- Nardozzi, J. D., Lott, K. and Cingolani, G. (2010). Phosphorylation meets nuclear import: a review. *Cell Communication and Signaling*, 8(1), 32. doi: 10.1186/1478-811x-8-32
- National Center for Biotechnology Information. (2002). Chapter 18, The Reference Sequence (RefSeq) Project. from National Library of Medicine (US), National Center for Biotechnology Information Available from [www.ncbi.nlm.nih.gov/books/NBK21091/](http://www.ncbi.nlm.nih.gov/books/NBK21091/)
- Nawaz, Z., Lonard, D. M., Dennis, A. P., Smith, C. L. and O'Malley, B. W. (1999). Proteasome-dependent degradation of the human estrogen receptor. *Proceedings of the National Academy of Sciences*, 96(5), 1858-1862. doi: 10.1073/pnas.96.5.1858
- NCBI Resource Coordinators. (2016). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, 44(Database issue), D7-D19. doi: 10.1093/nar/gkv1290
- Neal, J. A. and Meek, K. (2011). Choosing the right path: does DNA-PK help make the decision? *Mutat Res*, 711(1-2), 73-86. doi: 10.1016/j.mrfmmm.2011.02.010
- Negrini, S., Gorgoulis, V. G. and Halazonetis, T. D. (2010). Genomic instability - an evolving hallmark of cancer. *Nat Rev Mol Cell Biol*, 11(3), 220-228.
- Neph, S., Vierstra, J., Stergachis, A. B., Reynolds, A. P., Haugen, E., Vernot, B., *et al.* (2012). An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature*, 489(7414), 83-90. doi: 10.1038/nature11212
- Neri, F., Rapelli, S., Krepelova, A., Incarnato, D., Parlato, C., Basile, G., *et al.* (2017). Intragenic DNA methylation prevents spurious transcription initiation. *Nature*, 543(7643), 72-77. doi: 10.1038/nature21373
- Neumann, H., Hancock, S. M., Buning, R., Routh, A., Chapman, L., Somers, J., *et al.* (2009). A method for genetically installing site-specific acetylation in recombinant histones defines the effects of H3 K56 acetylation. *Mol Cell*, 36(1), 153-163. doi: 10.1016/j.molcel.2009.07.027
- Neve, R. M., Chin, K., Fridlyand, J., Yeh, J., Baehner, F. L., Fevr, T., *et al.* (2006). A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes. *Cancer Cell*, 10(6), 515-527. doi: 10.1016/j.ccr.2006.10.008
- Ng, R. K., Dean, W., Dawson, C., Lucifero, D., Madeja, Z., Reik, W., *et al.* (2008). Epigenetic restriction of embryonic cell lineage fate by methylation of Elf5. *Nat Cell Biol*, 10(11), 1280-1290. doi: 10.1038/ncb1786
- Nicholson, R. I., Hutcheson, I. R., Hiscox, S., Taylor, K. M. and Gee, J. M. W. (2009). Experimental Endocrine Resistance: Concepts and Strategies. In S. Hiscox, J. Gee, & R. I. Nicholson (Eds.), *Therapeutic Resistance to Anti-Hormonal Drugs in Breast Cancer: New Molecular Aspects and their Potential as Targets* (pp. 1-26). Dordrecht: Springer Netherlands.
- Nicholson, R. I., Staka, C., Boyns, F., Hutcheson, I. R. and Gee, J. M. W. (2004). Growth factor-driven mechanisms associated with resistance to estrogen deprivation in breast cancer: new opportunities for therapy. *Endocrine-Related Cancer*, 11(4), 623-641. doi: 10.1677/erc.1.00778
- Nie, Z., Hu, G., Wei, G., Cui, K., Yamane, A., Resch, W., *et al.* (2012). c-Myc is a universal amplifier of expressed genes in lymphocytes and embryonic stem cells. *Cell*, 151(1), 68-79. doi: 10.1016/j.cell.2012.08.033
- Niederriter, A. R., Varshney, A., Parker, S. C. and Martin, D. M. (2015). Super Enhancers in Cancers, Complex Disease, and Developmental Disorders. *Genes (Basel)*, 6(4), 1183-1200. doi: 10.3390/genes6041183
- Niesen, M. I., Osborne, A. R., Yang, H., Rastogi, S., Chellappan, S., Cheng, J. Q., *et al.* (2005). Activation of a methylated promoter mediated by a sequence-specific DNA-binding protein, RFX. *J Biol Chem*, 280(47), 38914-38922. doi: 10.1074/jbc.M504633200
- Nightingale, K. (2011). Histone Modifications and Their Role as Epigenetic Marks *Genome Organization and Function in the Cell Nucleus* (pp. 89-110): Wiley-VCH Verlag GmbH & Co. KGaA.
- Nik-Zainal, S., Davies, H., Staaf, J., Ramakrishna, M., Glodzik, D., Zou, X., *et al.* (2016). Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature*, 534(7605), 47-54. doi:



- 10.1038/nature17676
- Nock, A., Ascano, J. M., Jones, T., Barrero, M. J., Sugiyama, N., Tomita, M., *et al.* (2009). Identification of DNA-dependent protein kinase as a cofactor for the forkhead transcription factor FoxA2. *J Biol Chem*, 284(30), 19915-19926. doi: 10.1074/jbc.M109.016295
- Noguchi, T., Shibata, T., Fumoto, S., Uchida, Y., Mueller, W. and Takeno, S. (2002). DNA-PKcs expression in esophageal cancer as a predictor for chemoradiation therapeutic sensitivity. *Ann Surg Oncol*, 9(10), 1017-1022. doi: 10.1007/bf02574522
- O'Brate, A. and Giannakakou, P. (2003). The importance of p53 location: nuclear or cytoplasmic zip code? *Drug Resist Updat*, 6(6), 313-322.
- Oakes, S. R., Naylor, M. J., Asselin-Labat, M. L., Blazek, K. D., Gardiner-Garden, M., Hilton, H. N., *et al.* (2008). The Ets transcription factor Elf5 specifies mammary alveolar cell fate. *Genes Dev*, 22(5), 581-586. doi: 10.1101/gad.1614608
- Obenauer, J. C., Cantley, L. C. and Yaffe, M. B. (2003). Scansite 2.0: Proteome-wide prediction of cell signaling interactions using short sequence motifs. *Nucleic Acids Res*, 31(13), 3635-3641.
- Odom, D. T. (2011). Identification of Transcription Factor-DNA Interactions In Vivo. In T. R. Hughes (Ed.), *A Handbook of Transcription FactorsSubcellular Biochemistry* (Vol. 52, pp. 175-191): Springer. Retrieved from [www.ncbi.nlm.nih.gov/pubmed/21557083](http://www.ncbi.nlm.nih.gov/pubmed/21557083). doi: 10.1007/978-90-481-9069-0\_8
- Oettgen, P., Kas, K., Dube, A., Gu, X., Grall, F., Thamrongsak, U., *et al.* (1999). Characterization of ESE-2, a novel ESE-1-related Ets transcription factor that is restricted to glandular epithelium and differentiated keratinocytes. *J Biol Chem*, 274(41), 29439-29452.
- Oikawa, T. and Yamada, T. (2003). Molecular biology of the Ets family of transcription factors. *Gene*, 303, 11-34. doi: S0378111902011563 [pii]
- Oliva, R., Bazett-Jones, D. P., Locklear, L. and Dixon, G. H. (1990). Histone hyperacetylation can induce unfolding of the nucleosome core particle. *Nucleic Acids Res*, 18(9), 2739-2747.
- Ooi, S. K., Qiu, C., Bernstein, E., Li, K., Jia, D., Yang, Z., *et al.* (2007). DNMT3L connects unmethylated lysine 4 of histone H3 to de novo methylation of DNA. *Nature*, 448(7154), 714-717. doi: 10.1038/nature05987
- Org, T., Chignola, F., Hetenyi, C., Gaetani, M., Rebane, A., Liiv, I., *et al.* (2008). The autoimmune regulator PHD finger binds to non-methylated histone H3K4 to activate gene expression. *EMBO Rep*, 9(4), 370-376. doi: 10.1038/sj.embor.2008.11
- Pal, S., Gupta, R. and Davuluri, R. V. (2012). Alternative transcription and alternative splicing in cancer. *Pharmacol Ther*, 136(3), 283-294. doi: 10.1016/j.pharmthera.2012.08.005
- Panagopoulos, I., Gorunova, L., Davidson, B. and Heim, S. (2015). Novel TNS3-MAP3K3 and ZFPM2-ELF5 fusion genes identified by RNA sequencing in multicystic mesothelioma with t(7;17)(p12;q23) and t(8;11)(q23;p13). *Cancer Lett*, 357(2), 502-509. doi: 10.1016/j.canlet.2014.12.002
- Pankotai, T., Bonhomme, C., Chen, D. and Soutoglou, E. (2012). DNAPKcs-dependent arrest of RNA polymerase II transcription in the presence of DNA breaks. *Nat Struct Mol Biol*, 19(3), 276-282. doi: 10.1038/nsmb.2224
- Papamichos-Chronakis, M., Watanabe, S., Rando, O. J. and Peterson, C. L. (2011). Global regulation of H2A.Z localization by the INO80 chromatin-remodeling enzyme is essential for genome integrity. *Cell*, 144(2), 200-213. doi: 10.1016/j.cell.2010.12.021
- Pardee, K., Necakov, A. S. and Krause, H. (2011). Nuclear Receptors: Small Molecule Sensors that Coordinate Growth, Metabolism and Reproduction. In T. R. Hughes (Ed.), *A Handbook of Transcription FactorsSubcellular Biochemistry* (Vol. 52, pp. 123-153): Springer. Retrieved from [www.ncbi.nlm.nih.gov/pubmed/21557081](http://www.ncbi.nlm.nih.gov/pubmed/21557081). doi: 10.1007/978-90-481-9069-0\_6
- Park, S.-J., Gavrilova, O., Brown, A. L., Soto, J. E., Bremner, S., Kim, J., *et al.* (2017). DNA-PK Promotes the Mitochondrial, Metabolic, and Physical Decline that Occurs During Aging. *Cell Metabolism*, 25(5), 1135-1146.e1137. doi: 10.1016/j.cmet.2017.04.008
- Parker, B. S., Rautela, J. and Hertzog, P. J. (2016). Antitumour actions of interferons: implications for cancer therapy. *Nat Rev Cancer*, 16(3), 131-144. doi: 10.1038/nrc.2016.14
- Parker, J. S., Mullins, M., Cheang, M. C., Leung, S., Voduc, D., Vickery, T., *et al.* (2009). Supervised risk predictor of breast cancer based on intrinsic subtypes. *J Clin Oncol*, 27(8), 1160-1167. doi: 10.1200/JCO.2008.18.1370
- Parkinson, J., Lees-Miller, S. P. and Everett, R. D. (1999). Herpes simplex virus type 1 immediate-early protein vmw110 induces the proteasome-dependent degradation of the catalytic subunit of DNA-dependent protein kinase. *J Virol*, 73(1), 650-657.
- Patani, N. and Martin, L. A. (2014). Understanding response and resistance to oestrogen deprivation in ER-positive breast cancer. *Mol Cell Endocrinol*, 382(1), 683-694. doi: 10.1016/j.mce.2013.09.038
- Pencovich, N., Jaschek, R., Tanay, A. and Groner, Y. (2011). Dynamic combinatorial interactions of RUNX1 and cooperating partners regulates megakaryocytic differentiation in cell line models. *Blood*, 117(1), e1-14. doi: 10.1182/blood-2010-07-295113
- Perdigao-Henriques, R., Petrocca, F., Altschuler, G., Thomas, M. P., Le, M. T., Tan, S. M., *et al.* (2016). miR-200 promotes the mesenchymal to epithelial transition by suppressing multiple members of the Zeb2 and Snail1 transcriptional repressor complexes. *Oncogene*, 35(2), 158-172. doi: 10.1038/nc.2015.69
- Pereira, B., Chin, S. F., Rueda, O. M., Volland, H. K., Provenzano, E., Bardwell, H. A., *et al.* (2016). The somatic mutation profiles of 2,433 breast cancers refines their genomic and transcriptomic landscapes. *Nat Commun*, 7, 11479. doi: 10.1038/ncomms11479
- Perkins, N. D. (2007). Integrating cell-signalling pathways with NF-KB and IKK function. *Nat Rev Mol Cell Biol*, 8(1), 49-62.
- Perou, C. M. and Borresen-Dale, A. L. (2011). Systems biology and genomics of breast cancer. *Cold Spring Harb Perspect Biol*, 3(2). doi: 10.1101/cshperspect.a003293
- Perou, C. M., Sorlie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Rees, C. A., *et al.* (2000). Molecular portraits of human breast tumours. *Nature*, 406(6797), 747-752. doi: 10.1038/35021093
- Peters, N. E., Ferguson, B. J., Mazzon, M., Fahy, A. S., Krysztofinska, E., Arribas-Bosacoma, R., *et al.* (2013). A Mechanism for the Inhibition of DNA-PK-Mediated DNA Sensing by a Virus. *PLOS Pathogens*, 9(10), e1003649. doi: 10.1371/journal.ppat.1003649
- Peterson, T. J., Karmakar, S., Pace, M. C., Gao, T. and Smith, C. L. (2007). The silencing mediator of retinoic acid and thyroid hormone receptor (SMRT) corepressor is required for full estrogen receptor alpha transcriptional activity. *Mol Cell Biol*, 27(17), 5933-5948. doi: 10.1128/MCB.00237-07

- Pfaffl, M. W. (2001). A new mathematical model for relative quantification in real-time RT-PCR. *Nucleic Acids Res*, 29(9), e45.
- Piggin, C. L., Roden, D. L., Gallego-Ortega, D., Lee, H. J., Oakes, S. R. and Ormandy, C. J. (2016). ELF5 isoform expression is tissue-specific and significantly altered in cancer. *Breast Cancer Res*, 18(1), 4. doi: 10.1186/s13058-015-0666-0
- Pott, S. and Lieb, J. D. (2015). What are super-enhancers? *Nat Genet*, 47(1), 8-12. doi: 10.1038/ng.3167
- Pradeepa, M. M., Grimes, G. R., Kumar, Y., Olley, G., Taylor, G. C., Schneider, R., *et al.* (2016). Histone H3 globular domain acetylation identifies a new class of enhancers. *Nat Genet*, 48(6), 681-686. doi: 10.1038/ng.3550
- Prat, A., Carey, L. A., Adamo, B., Vidal, M., Tabernero, J., Cortés, J., *et al.* (2014). Molecular Features and Survival Outcomes of the Intrinsic Subtypes Within HER2-Positive Breast Cancer. *JNCI: Journal of the National Cancer Institute*, 106(8), dju152-dju152. doi: 10.1093/jnci/dju152
- Prat, A., Karginova, O., Parker, J. S., Fan, C., He, X., Bixby, L., *et al.* (2013). Characterization of cell lines derived from breast cancers and normal mammary tissues for the study of the intrinsic molecular subtypes. *Breast Cancer Res Treat*, 142(2), 237-255. doi: 10.1007/s10549-013-2743-3
- Prat, A., Parker, J. S., Karginova, O., Fan, C., Livasy, C., Herschkowitz, J. I., *et al.* (2010). Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer. *Breast Cancer Res*, 12(5), R68. doi: 10.1186/bcr2635
- Prat, A. and Perou, C. M. (2009). Mammary development meets cancer genomics. *Nat Med*, 15(8), 842-844. doi: 10.1038/nm0809-842
- Prat, A. and Perou, C. M. (2011). Deconstructing the molecular portraits of breast cancer. *Mol Oncol*, 5(1), 5-23. doi: 10.1016/j.molonc.2010.11.003
- Prat, A., Pineda, E., Adamo, B., Galvan, P., Fernandez, A., Gaba, L., *et al.* (2015). Clinical implications of the intrinsic molecular subtypes of breast cancer. *Breast*, 24 Suppl 2, S26-35. doi: 10.1016/j.breast.2015.07.008
- Prescott, J. D., Koto, K. S., Singh, M. and Gutierrez-Hartmann, A. (2004). The ETS transcription factor ESE-1 transforms MCF-12A human mammary epithelial cells via a novel cytoplasmic mechanism. *Mol Cell Biol*, 24(12), 5548-5564. doi: 10.1128/MCB.24.12.5548-5564.2004
- Prescott, J. D., Pocobutt, J. M., Tentler, J. J., Walker, D. M. and Gutierrez-Hartmann, A. (2011). Mapping of ESE-1 subdomains required to initiate mammary epithelial cell transformation via a cytoplasmic mechanism. *Mol Cancer*, 10, 103. doi: 10.1186/1476-4598-10-103
- Prokhorchouk, A., Hendrich, B., Jorgensen, H., Ruzov, A., Wilm, M., Georgiev, G., *et al.* (2001). The p120 catenin partner Kaiso is a DNA methylation-dependent transcriptional repressor. *Genes Dev*, 15(13), 1613-1618. doi: 10.1101/gad.198501
- Pufall, M. A., Lee, G. M., Nelson, M. L., Kang, H. S., Velyvis, A., Kay, L. E., *et al.* (2005). Variable control of Ets-1 DNA binding by multiple phosphates in an unstructured region. *Science*, 309(5731), 142-145. doi: 10.1126/science.1111915
- Pyun, B. J., Seo, H. R., Lee, H. J., Jin, Y. B., Kim, E. J., Kim, N. H., *et al.* (2013). Mutual regulation between DNA-PKcs and Snail1 leads to increased genomic instability and aggressive tumor characteristics. *Cell Death Dis*, 4, e517. doi: 10.1038/cddis.2013.43
- Quinlan, A. R. and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6), 841-842. doi: 10.1093/bioinformatics/btq033
- Quinn, J. J. and Chang, H. Y. (2016). Unique features of long non-coding RNA biogenesis and function. *Nat Rev Genet*, 17(1), 47-62. doi: 10.1038/nrg.2015.10
- Radhakrishnan, A., Raju, R., Tuladhar, N., Subbannayya, T., Thomas, J. K., Goel, R., *et al.* (2012). A pathway map of prolactin signaling. *J Cell Commun Signal*, 6(3), 169-173. doi: 10.1007/s12079-012-0168-0
- Rahim, G., Araud, T., Jaquier-Gubler, P. and Curran, J. (2012). Alternative splicing within the elk-1 5' untranslated region serves to modulate initiation events downstream of the highly conserved upstream open reading frame 2. *Mol Cell Biol*, 32(9), 1745-1756. doi: 10.1128/MCB.06751-11
- Rahim, S., Beauchamp, E. M., Kong, Y., Brown, M. L., Toretsky, J. A. and Üren, A. (2011). YK-4-279 Inhibits ERG and ETV1 Mediated Prostate Cancer Cell Invasion. *PLoS One*, 6(4), e19343. doi: 10.1371/journal.pone.0019343
- Rahl, P. B., Lin, C. Y., Seila, A. C., Flynn, R. A., McCuine, S., Burge, C. B., *et al.* (2010). c-Myc regulates transcriptional pause release. *Cell*, 141(3), 432-445. doi: 10.1016/j.cell.2010.03.030
- Ramirez, F., Dundar, F., Diehl, S., Gruning, B. A. and Manke, T. (2014). deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Res*, 42(Web Server issue), W187-191. doi: 10.1093/nar/gku365
- Rando, O. J. (2012). Combinatorial complexity in chromatin structure and function: revisiting the histone code. *Curr Opin Genet Dev*, 22(2), 148-155. doi: 10.1016/j.gde.2012.02.013
- Ratman, D., Vanden Berghe, W., Dejager, L., Libert, C., Tavernier, J., Beck, I. M., *et al.* (2013). How glucocorticoid receptors modulate the activity of other transcription factors: a scope beyond tethering. *Mol Cell Endocrinol*, 380(1-2), 41-54. doi: 10.1016/j.mce.2012.12.014
- Reich, M., Liefeld, T., Gould, J., Lerner, J., Tamayo, P. and Mesirov, J. P. (2006). GenePattern 2.0. *Nat Genet*, 38(5), 500-501. doi: 10.1038/ng0506-500
- Rekhtman, N., Radparvar, F., Evans, T. and Skoultschi, A. I. (1999). Direct interaction of hematopoietic transcription factors PU.1 and GATA-1: functional antagonism in erythroid cells. *Genes Dev*, 13(11), 1398-1411.
- Renda, M., Baglivo, I., Burgess-Beusse, B., Esposito, S., Fattorusso, R., Felsenfeld, G., *et al.* (2007). Critical DNA binding interactions of the insulator protein CTCF: a small number of zinc fingers mediate strong binding, and a single finger-DNA interaction controls binding at imprinted loci. *J Biol Chem*, 282(46), 33336-33345. doi: 10.1074/jbc.M706213200
- Rice, J. C. and Allis, C. D. (2001). Histone methylation versus histone acetylation: new insights into epigenetic regulation. *Curr Opin Cell Biol*, 13(3), 263-273.
- Rigas, B., Borgo, S., Elhosseiny, A., Balatsos, V., Manika, Z., Shinya, H., *et al.* (2001). Decreased expression of DNA-dependent protein kinase, a DNA repair protein, during human colon carcinogenesis. *Cancer Res*, 61(23), 8381-8384.
- Riising, E. M., Comet, I., Leblanc, B., Wu, X., Johansen, J. V. and Helin, K. (2014). Gene silencing triggers polycomb repressive complex 2 recruitment to CpG islands genome wide. *Mol Cell*, 55(3), 347-360. doi: 10.1016/j.molcel.2014.06.005
- Rishi, V., Bhattacharya, P., Chatterjee, R., Rozenberg, J., Zhao, J., Glass, K., *et al.* (2010). CpG methylation of half-CRE sequences creates C/EBPalpha binding sites that activate some tissue-specific genes. *Proc Natl Acad*

- Sci U S A*, 107(47), 20311-20316. doi: 10.1073/pnas.1008688107
- Risinger, J. I., Maxwell, G. L., Chandramouli, G. V., Jazaeri, A., Aprelikova, O., Patterson, T., *et al.* (2003). Microarray analysis reveals distinct gene expression profiles among different histologic types of endometrial cancer. *Cancer Res*, 63(1), 6-11.
- Robinson, J. L. and Carroll, J. S. (2012). FoxA1 is a key mediator of hormonal response in breast and prostate cancer. *Front Endocrinol (Lausanne)*, 3, 68. doi: 10.3389/fendo.2012.00068
- Robinson, J. T., Thorvaldsdottir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., *et al.* (2011). Integrative genomics viewer. *Nat Biotech*, 29(1), 24-26. doi: 10.1038/nbt.1754
- Robinson, M. D., McCarthy, D. J. and Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1), 139-140. doi: 10.1093/bioinformatics/btp616
- Robinson, M. D. and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol*, 11(3), R25. doi: 10.1186/gb-2010-11-3-r25
- Robinson, M. D. and Smyth, G. K. (2007). Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics*, 23(21), 2881-2887. doi: 10.1093/bioinformatics/btm453
- Robinson, M. D. and Smyth, G. K. (2008). Small-sample estimation of negative binomial dispersion, with applications to SAGE data. *Biostatistics*, 9(2), 321-332. doi: 10.1093/biostatistics/kxm030
- Roos, W. P., Thomas, A. D. and Kaina, B. (2016). DNA damage and the balance between survival and death in cancer biology. *Nat Rev Cancer*, 16(1), 20-33. doi: 10.1038/nrc.2015.2
- Ross-Innes, C. S., Stark, R., Teschendorff, A. E., Holmes, K. A., Ali, H. R., Dunning, M. J., *et al.* (2012). Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. *Nature*, 481(7381), 389-393. doi: 10.1038/nature10730
- Rothbart, S. B. and Strahl, B. D. (2014). Interpreting the language of histone and DNA modifications. *Biochim Biophys Acta*, 1839(8), 627-643. doi: 10.1016/j.bbagr.2014.03.001
- Ruthenburg, A. J., Li, H., Milne, T. A., Dewell, S., McGinty, R. K., Yuen, M., *et al.* (2011). Recognition of a mononucleosomal histone modification pattern by BPTF via multivalent interactions. *Cell*, 145(5), 692-706. doi: 10.1016/j.cell.2011.03.053
- Sanders, D. A., Ross-Innes, C. S., Beraldi, D., Carroll, J. S. and Balasubramanian, S. (2013). Genome-wide mapping of FOXM1 binding reveals co-binding with estrogen receptor alpha in breast cancer cells. *Genome Biol*, 14(1), R6. doi: 10.1186/gb-2013-14-1-r6
- Santen, R. J., Song, R. X., Zhang, Z., Yue, W. and Kumar, R. (2004). Adaptive Hypersensitivity to Estrogen. *Mechanism for Sequential Responses to Hormonal Therapy in Breast Cancer*, 10(1), 337s-345s. doi: 10.1158/1078-0432.ccr-031207
- Sanyal, A., Lajoie, B. R., Jain, G. and Dekker, J. (2012). The long-range interaction landscape of gene promoters. *Nature*, 489(7414), 109-113. doi: 10.1038/nature11279
- Sari, I. N., Yang, Y. G., Phi, L. T., Kim, H., Baek, M. J., Jeong, D., *et al.* (2016). Interferon-induced transmembrane protein 1 (IFITM1) is required for the progression of colorectal cancer. *Oncotarget*, 7(52), 86039-86050. doi: 10.18632/oncotarget.13325
- Sartorius, C. A., Takimoto, G. S., Richer, J. K., Tung, L. and Horwitz, K. B. (2000). Association of the Ku autoantigen/DNA-dependent protein kinase holoenzyme and poly(ADP-ribose) polymerase with the DNA binding domain of progesterone receptors. *J Mol Endocrinol*, 24(2), 165-182.
- Sato, N., Kondo, M. and Arai, K.-i. (2006). The orphan nuclear receptor GCNF recruits DNA methyltransferase for Oct-3/4 silencing. *Biochemical and Biophysical Research Communications*, 344(3), 845-851. doi: 10.1016/j.bbrc.2006.04.007
- Schild-Poulter, C., Pope, L., Giffin, W., Kochan, J. C., Ngsee, J. K., Traykova-Andonova, M., *et al.* (2001). The binding of Ku antigen to homeodomain proteins promotes their phosphorylation by DNA-dependent protein kinase. *J Biol Chem*, 276(20), 16848-16856. doi: 10.1074/jbc.M100768200
- Schild-Poulter, C., Shih, A., Tantin, D., Yarymowich, N. C., Soubeyrand, S., Sharp, P. A., *et al.* (2007). DNA-PK phosphorylation sites on Oct-1 promote cell survival following DNA damage. *Oncogene*, 26(27), 3980-3988. doi: 10.1038/sj.onc.1210165
- Schild-Poulter, C., Shih, A., Yarymowich, N. C. and Haché, R. J. G. (2003). Down-Regulation of Histone H2B by DNA-Dependent Protein Kinase in Response to DNA Damage through Modulation of Octamer Transcription Factor 1. *Cancer Research*, 63(21), 7197-7205.
- Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., *et al.* (2012). Fiji: an open-source platform for biological-image analysis. *Nat Methods*, 9(7), 676-682. doi: 10.1038/nmeth.2019
- Schmidl, C., Klug, M., Boeld, T. J., Andreesen, R., Hoffmann, P., Edinger, M., *et al.* (2009). Lineage-specific DNA methylation in T cells correlates with histone methylation and enhancer activity. *Genome Res*, 19(7), 1165-1174. doi: 10.1101/gr.091470.109
- Schoenberg, D. R. and Maquat, L. E. (2012). Regulation of cytoplasmic mRNA decay. *Nat Rev Genet*, 13(4), 246-259.
- Schwartz-Roberts, J. L., Cook, K. L., Chen, C., Shajahan-Haq, A. N., Axelrod, M., Warri, A., *et al.* (2015). Interferon regulatory factor-1 signaling regulates the switch between autophagy and apoptosis to determine breast cancer cell fate. *Cancer Res*, 75(6), 1046-1055. doi: 10.1158/0008-5472.CAN-14-1851
- Schwartz, J. L., Shajahan, A. N. and Clarke, R. (2011). The Role of Interferon Regulatory Factor-1 (IRF1) in Overcoming Antiestrogen Resistance in the Treatment of Breast Cancer. *Int J Breast Cancer*, 2011, 912102. doi: 10.4061/2011/912102
- Scully, K. M., Jacobson, E. M., Jepsen, K., Lunyak, V., Viadiu, H., Carriere, C., *et al.* (2000). Allosteric effects of Pit-1 DNA sites on long-term repression in cell type specification. *Science*, 290(5494), 1127-1131.
- Sebestyen, E., Zawisza, M. and Eyra, E. (2015). Detection of recurrent alternative splicing switches in tumor samples reveals novel signatures of cancer. *Nucleic Acids Res*, 43(3), 1345-1356. doi: 10.1093/nar/gku1392
- Seeler, J.-S. and Dejean, A. (2017). SUMO and the robustness of cancer. *Nat Rev Cancer*, 17(3), 184-197. doi: 10.1038/nrc.2016.143
- Seki, Y., Suico, M. A., Uto, A., Hisatsune, A., Shuto, T., Isohama, Y., *et al.* (2002). The ETS Transcription Factor MEF Is a Candidate Tumor Suppressor Gene on the X Chromosome. *Cancer Research*, 62(22), 6579-6586.
- Sekiya, T., Muthurajan, U. M., Luger, K., Tulin, A. V. and Zaret, K. S. (2009). Nucleosome-binding affinity as a primary determinant of the nuclear mobility of the pioneer transcription factor FoxA. *Genes Dev*, 23(7), 804-809. doi: 10.1101/gad.1775509
- Sekiya, T. and Zaret, K. S. (2007). Repression by Groucho/TLE/Grg proteins: genomic site recruitment generates

- compacted chromatin in vitro and impairs activator binding in vivo. *Mol Cell*, 28(2), 291-303. doi: 10.1016/j.molcel.2007.10.002
- SEQC/MAQC-III Consortium. (2014). A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. *Nat Biotech*, 32(9), 903-914. doi: 10.1038/nbt.2957
- Serandour, A. A., Avner, S., Percevault, F., Demay, F., Bizot, M., Lucchetti-Miganeh, C., *et al.* (2011). Epigenetic switch involved in activation of pioneer factor FOXA1-dependent enhancers. *Genome Res*, 21(4), 555-565. doi: 10.1101/gr.111534.110
- Seshasayee, A. S., Sivaraman, K. and Luscombe, N. M. (2011). An overview of prokaryotic transcription factors : a summary of function and occurrence in bacterial genomes. *Subcell Biochem*, 52, 7-23. doi: 10.1007/978-90-481-9069-0\_2
- Seth, A. and Watson, D. K. (2005). ETS transcription factors and their emerging roles in human cancer. *Eur J Cancer*, 41(16), 2462-2478. doi: 10.1016/j.ejca.2005.08.013
- Seto, E. and Yoshida, M. (2014). Erasers of histone acetylation: the histone deacetylase enzymes. *Cold Spring Harb Perspect Biol*, 6(4), a018713. doi: 10.1101/cshperspect.a018713
- Shackleton, M., Vaillant, F., Simpson, K. J., Stingl, J., Smyth, G. K., Asselin-Labat, M.-L., *et al.* (2006). Generation of a functional mammary gland from a single stem cell. *Nature*, 439(7072), 84-88. doi: 10.1038/nature04372
- Shah, S. P., Roth, A., Goya, R., Oloumi, A., Ha, G., Zhao, Y., *et al.* (2012). The clonal and mutational evolution spectrum of primary triple-negative breast cancers. *Nature*, 486(7403), 395-399. doi: 10.1038/nature10933
- Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., *et al.* (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*, 13(11), 2498-2504. doi: 10.1101/gr.1239303
- Shao, C.-J., Fu, J., Shi, H.-L., Mu, Y.-G. and Chen, Z.-P. (2008). Activities of DNA-PK and Ku86, but not Ku70, may predict sensitivity to cisplatin in human gliomas. *Journal of Neuro-Oncology*, 89(1), 27. doi: 10.1007/s11060-008-9592-7
- Sharif, H., Li, Y., Dong, Y., Dong, L., Wang, W. L., Mao, Y., *et al.* (2017). Cryo-EM structure of the DNA-PK holoenzyme. *Proc Natl Acad Sci U S A*. doi: 10.1073/pnas.1707386114
- Sharma, K., D'Souza, Rochelle C. J., Tyanova, S., Schaab, C., Wiśniewski, Jacek R., Cox, J., *et al.* (2014). Ultradeep Human Phosphoproteome Reveals a Distinct Regulatory Nature of Tyr and Ser/Thr-Based Signaling. *Cell Reports*, 8(5), 1583-1594. doi: 10.1016/j.celrep.2014.07.036
- Sharrocks, A. D. (2001). The ETS-domain transcription factor family. *Nat Rev Mol Cell Biol*, 2(11), 827-837. doi: 10.1038/35099076
- Sharrocks, A. D., Yang, S.-H. and Galanis, A. (2000). Docking domains and substrate-specificity determination for MAP kinases. *Trends Biochem Sci*, 25(9), 448-453. doi: 10.1016/S0968-0004(00)01627-3
- Shenoy, A. and Belloch, R. H. (2014). Regulation of microRNA function in somatic stem cell proliferation and differentiation. *Nat Rev Mol Cell Biol*, 15(9), 565-576. doi: 10.1038/nrm3854
- Sherwood, R. I., Hashimoto, T., O'Donnell, C. W., Lewis, S., Barkal, A. A., van Hoff, J. P., *et al.* (2014). Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape. *Nat Biotechnol*, 32(2), 171-178. doi: 10.1038/nbt.2798
- Shi, Y., Lan, F., Matson, C., Mulligan, P., Whetstone, J. R., Cole, P. A., *et al.* (2004). Histone demethylation mediated by the nuclear amine oxidase homolog LSD1. *Cell*, 119(7), 941-953. doi: 10.1016/j.cell.2004.12.012
- Shimomura, A., Takasaki, A., Nomura, R., Hayashi, N. and Senda, T. (2013). Identification of DNA-dependent protein kinase catalytic subunit as a novel interaction partner of lymphocyte enhancer factor 1. *Med Mol Morphol*, 46(1), 14-19. doi: 10.1007/s00795-012-0002-z
- Shin, H. Y., Willi, M., Yoo, K. H., Zeng, X., Wang, C., Metser, G., *et al.* (2016). Hierarchy within the mammary STAT5-driven Wap super-enhancer. *Nat Genet*, 48(8), 904-911. doi: 10.1038/ng.3606
- Shintani, S., Mihara, M., Li, C., Nakahara, Y., Hino, S., Nakashiro, K., *et al.* (2003). Up-regulation of DNA-dependent protein kinase correlates with radiation resistance in oral squamous cell carcinoma. *Cancer Sci*, 94(10), 894-900.
- Shlyueva, D., Stampfel, G. and Stark, A. (2014). Transcriptional enhancers: from properties to genome-wide predictions. *Nat Rev Genet*, 15(4), 272-286. doi: 10.1038/nrg3682
- Shrivastava, T., Mino, K., Babayeva, N. D., Baranovskaya, O. I., Rizzino, A. and Tahirov, T. H. (2014). Structural basis of Ets1 activation by Runx1. *Leukemia*, 28(10), 2040-2048. doi: 10.1038/leu.2014.111
- Shukla, S., Kavak, E., Gregory, M., Imashimizu, M., Shutinoski, B., Kashlev, M., *et al.* (2011). CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. *Nature*, 479(7371), 74-79. doi: 10.1038/nature10442
- Sibanda, B. L., Chirgadze, D. Y., Ascher, D. B. and Blundell, T. L. (2017). DNA-PKcs structure suggests an allosteric mechanism modulating DNA double-strand break repair. *Science*, 355(6324), 520-524. doi: 10.1126/science.aak9654
- Siersbaek, R., Rabiee, A., Nielsen, R., Sidoli, S., Traynor, S., Loft, A., *et al.* (2014). Transcription factor cooperativity in early adipogenic hotspots and super-enhancers. *Cell Rep*, 7(5), 1443-1455. doi: 10.1016/j.celrep.2014.04.042
- Sigma-Aldrich. (2013). *User Guide Duolink In Situ - Fluorescence* St Louis, Missouri: Sigma-Aldrich.
- Sikora-Wohlfeld, W., Ackermann, M., Christodoulou, E. G., Singaravelu, K. and Beyer, A. (2013). Assessing Computational Methods for Transcription Factor Target Gene Identification Based on ChIP-seq Data. *PLoS Comput Biol*, 9(11), e1003342. doi: 10.1371/journal.pcbi.1003342
- Sims, R. J. and Reinberg, D. (2008). Is there a code embedded in proteins that is based on post-translational modifications? *Nat Rev Mol Cell Biol*, 9(10), 815-820.
- Sizemore, G. M., Pitarresi, J. R., Balakrishnan, S. and Ostrowski, M. C. (2017). The ETS family of oncogenic transcription factors in solid tumours. *Nat Rev Cancer, advance online publication*. doi: 10.1038/nrc.2017.20
- Skibinski, A. and Kuperwasser, C. (2015). The origin of breast tumor heterogeneity. *Oncogene*, 34(42), 5309-5316. doi: 10.1038/onc.2014.475
- Slattery, M., Zhou, T., Yang, L., Dantas Machado, A. C., Gordan, R. and Rohs, R. (2014). Absence of a simple code: how transcription factors read the genome. *Trends Biochem Sci*, 39(9), 381-399. doi: 10.1016/j.tibs.2014.07.002
- Smith, G. C. and Jackson, S. P. (1999). The DNA-dependent protein kinase. *Genes Dev*, 13(8), 916-934.
- Smolle, M., Venkatesh, S., Gogol, M. M., Li, H., Zhang, Y., Florens, L., *et al.* (2012). Chromatin remodelers Isw1 and Chd1 maintain chromatin structure during transcription by preventing histone exchange. *Nat Struct Mol Biol*,

- 19(9), 884-892. doi: 10.1038/nsmb.2312
- Smyth, G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol*, 3, Article3. doi: 10.2202/1544-6115.1027
- Soderlund Leifler, K., Queseth, S., Fornander, T. and Askmalm, M. S. (2010). Low expression of Ku70/80, but high expression of DNA-PKcs, predict good response to radiotherapy in early breast cancer. *Int J Oncol*, 37(6), 1547-1554.
- Sodir, N. M. and Evan, G. I. (2009). Nursing some sense out of Myc. *Journal of Biology*, 8(8), 77. doi: 10.1186/jbiol181
- Someya, M., Sakata, K.-i., Matsumoto, Y., Yamamoto, H., Monobe, M., Ikeda, H., *et al.* (2005). The association of DNA-dependent protein kinase activity with chromosomal instability and risk of cancer. *Carcinogenesis*, 27(1), 117-122. doi: 10.1093/carcin/bgi175
- Sorlie, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., *et al.* (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci U S A*, 98(19), 10869-10874. doi: 10.1073/pnas.191367098
- Soufi, A., Garcia, M. F., Jaroszewicz, A., Osman, N., Pellegrini, M. and Zaret, K. S. (2015). Pioneer transcription factors target partial DNA motifs on nucleosomes to initiate reprogramming. *Cell*, 161(3), 555-568. doi: 10.1016/j.cell.2015.03.017
- Spitz, F. and Furlong, E. E. (2012). Transcription factors: from enhancer binding to developmental control. *Nat Rev Genet*, 13(9), 613-626. doi: 10.1038/nrg3207
- Stadler, M. B., Murr, R., Burger, L., Ivanek, R., Lienert, F., Scholer, A., *et al.* (2011). DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature*, 480(7378), 490-495. doi: 10.1038/nature10716
- Staka, C. M., Nicholson, R. I. and Gee, J. M. (2005). Acquired resistance to oestrogen deprivation: role for growth factor signalling kinases/oestrogen receptor cross-talk revealed in new MCF-7X model. *Endocr Relat Cancer*, 12 Suppl 1, S85-97. doi: 10.1677/erc.1.01006
- Stellato, C., Porreca, I., Cuomo, D., Tarallo, R., Nassa, G. and Ambrosino, C. (2016). The "busy life" of unliganded estrogen receptors. *Proteomics*, 16(2), 288-300. doi: 10.1002/pmic.201500261
- Stingl, J., Eirew, P., Ricketson, I., Shackleton, M., Vaillant, F., Choi, D., *et al.* (2006). Purification and unique properties of mammary epithelial stem cells. *Nature*, 439(7079), 993-997. doi: 10.1038/nature04496
- Stoelzle, T., Schwarb, P., Trumpf, A. and Hynes, N. E. (2009). c-Myc affects mRNA translation, cell proliferation and progenitor cell function in the mammary gland. *BMC Biology*, 7(1), 63. doi: 10.1186/1741-7007-7-63
- Stopka, T., Amanatullah, D. F., Papetti, M. and Skoultschi, A. I. (2005). PU.1 inhibits the erythroid program by binding to GATA-1 on DNA and creating a repressive chromatin structure. *EMBO J*, 24(21), 3712-3723. doi: 10.1038/sj.emboj.7600834
- Strahl, B. D. and Allis, C. D. (2000). The language of covalent histone modifications. *Nature*, 403(6765), 41-45. doi: 10.1038/47412
- Strahl, B. D., Grant, P. A., Briggs, S. D., Sun, Z. W., Bone, J. R., Caldwell, J. A., *et al.* (2002). Set2 is a nucleosomal histone H3-selective methyltransferase that mediates transcriptional repression. *Mol Cell Biol*, 22(5), 1298-1306.
- Stricker, S. H., Koferle, A. and Beck, S. (2017). From profiles to function in epigenomics. *Nat Rev Genet*, 18(1), 51-66. doi: 10.1038/nrg.2016.138
- Struhl, K. and Segal, E. (2013). Determinants of nucleosome positioning. *Nat Struct Mol Biol*, 20(3), 267-273. doi: 10.1038/nsmb.2506
- Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., *et al.* (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*, 102(43), 15545-15550. doi: 10.1073/pnas.0506580102
- Sun, G., Yang, L., Dong, C., Ma, B., Shan, M. and Ma, B. (2017). PRKDC regulates chemosensitivity and is a potential prognostic and predictive marker of response to adjuvant chemotherapy in breast cancer patients. *Oncol Rep*, 37(6), 3536-3542. doi: 10.3892/or.2017.5634
- Sun, S., Cheng, S., Zhu, Y., Zhang, P., Liu, N., Xu, T., *et al.* (2016). Identification of PRKDC (Protein Kinase, DNA-Activated, Catalytic Polypeptide) as an essential gene for colorectal cancer (CRCs) cells. *Gene*, 584(1), 90-96. doi: 10.1016/j.gene.2016.03.020
- Surget, S., Khoury, M. P. and Bourdon, J. C. (2013). Uncovering the role of p53 splice variants in human malignancy: a clinical perspective. *Onco Targets Ther*, 7, 57-68. doi: 10.2147/OTT.S53876
- Surova, O. and Zhivotovsky, B. (2013). Various modes of cell death induced by DNA damage. *Oncogene*, 32(33), 3789-3797. doi: 10.1038/onc.2012.556
- Sveen, A., Johannessen, B., Teixeira, M. R., Lothe, R. A. and Skotheim, R. I. (2014). Transcriptome instability as a molecular pan-cancer characteristic of carcinomas. *BMC Genomics*, 15, 672. doi: 10.1186/1471-2164-15-672
- Swinstead, E. E., Miranda, T. B., Paakinaho, V., Baek, S., Goldstein, I., Hawkins, M., *et al.* (2016a). Steroid Receptors Reprogram FoxA1 Occupancy through Dynamic Chromatin Transitions. *Cell*, 165(3), 593-605. doi: 10.1016/j.cell.2016.02.067
- Swinstead, E. E., Paakinaho, V., Presman, D. M. and Hager, G. L. (2016b). Pioneer factors and ATP-dependent chromatin remodeling factors interact dynamically: A new perspective: Multiple transcription factors can effect chromatin pioneer functions through dynamic interactions with ATP-dependent chromatin remodeling factors. *Bioessays*, 38(11), 1150-1157. doi: 10.1002/bies.201600137
- Taberlay, P. C., Satham, A. L., Kelly, T. K., Clark, S. J. and Jones, P. A. (2014). Reconfiguration of nucleosome-depleted regions at distal regulatory elements accompanies DNA methylation of enhancers and insulators in cancer. *Genome Res*, 24(9), 1421-1432. doi: 10.1101/gr.163485.113
- Tahiliani, M., Koh, K. P., Shen, Y., Pastor, W. A., Bandukwala, H., Brudno, Y., *et al.* (2009). Conversion of 5-Methylcytosine to 5-Hydroxymethylcytosine in Mammalian DNA by MLL Partner TET1. *Science*, 324(5929), 930-935.
- Takahashi, K., Tanabe, K., Ohnuki, M., Narita, M., Ichisaka, T., Tomoda, K., *et al.* (2007). Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell*, 131(5), 861-872. doi: 10.1016/j.cell.2007.11.019
- Takaku, M., Grimm, S. A., Shimbo, T., Perera, L., Menafra, R., Stunnenberg, H. G., *et al.* (2016). GATA3-dependent cellular reprogramming requires activation-domain dependent recruitment of a chromatin remodeler. *Genome Biol*, 17, 36. doi: 10.1186/s13059-016-0897-0
- Talbert, P. B. and Henikoff, S. (2017). Histone variants on the move: substrates for chromatin dynamics. *Nat Rev Mol Cell Biol*, 18(2), 115-126. doi: 10.1038/nrm.2016.148

- Tamura, R., Yoshihara, K., Yamawaki, K., Suda, K., Ishiguro, T., Adachi, S., *et al.* (2015). Novel kinase fusion transcripts found in endometrial cancer. *Sci Rep*, 5, 18657. doi: 10.1038/srep18657
- Tenen, D. G. (2003). Disruption of differentiation in human cancer: AML shows the way. *Nat Rev Cancer*, 3(2), 89-101. doi: 10.1038/nrc989
- Tessarz, P. and Kouzarides, T. (2014). Histone core modifications regulating nucleosome structure and dynamics. *Nat Rev Mol Cell Biol*, 15(11), 703-708. doi: 10.1038/nrm3890
- Teves, S. S., Weber, C. M. and Henikoff, S. (2014). Transcribing through the nucleosome. *Trends Biochem Sci*, 39(12), 577-586. doi: 10.1016/j.tibs.2014.10.004
- The Uniprot Consortium. (2017). UniProt: the universal protein knowledgebase. *Nucleic Acids Research*, 45(D1), D158-D169. doi: 10.1093/nar/gkw1099
- Thijssen, R., Ter Burg, J., Garrick, B., van Bochove, G. G., Brown, J. R., Fernandes, S. M., *et al.* (2016). Dual TORK/DNA-PK inhibition blocks critical signaling pathways in chronic lymphocytic leukemia. *Blood*, 128(4), 574-583. doi: 10.1182/blood-2016-02-700328
- Thomson, J. P., Skene, P. J., Selfridge, J., Clouaire, T., Guy, J., Webb, S., *et al.* (2010). CpG islands influence chromatin structure via the CpG-binding protein Cfp1. *Nature*, 464(7291), 1082-1086. doi: 10.1038/nature08924
- Thorvaldsdóttir, H., Robinson, J. T. and Mesirov, J. P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform*, 14(2), 178-192. doi: 10.1093/bib/bbs017
- Thurman, R. E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M. T., Haugen, E., *et al.* (2012). The accessible chromatin landscape of the human genome. *Nature*, 489(7414), 75-82. doi: 10.1038/nature11232
- Tian, B. and Manley, J. L. (2013). Alternative cleavage and polyadenylation: the long and short of it. *Trends Biochem Sci*, 38(6), 312-320. doi: 10.1016/j.tibs.2013.03.005
- Tian, B. and Manley, J. L. (2017). Alternative polyadenylation of mRNA precursors. *Nat Rev Mol Cell Biol*, 18(1), 18-30. doi: 10.1038/nrm.2016.116
- Tillo, D., Kaplan, N., Moore, I. K., Fondufe-Mittendorf, Y., Gossett, A. J., Field, Y., *et al.* (2010). High nucleosome occupancy is encoded at human regulatory sequences. *PLoS One*, 5(2), e9129. doi: 10.1371/journal.pone.0009129
- Todeschini, A. L., Georges, A. and Veltia, R. A. (2014). Transcription factors: specific DNA binding and specific gene regulation. *Trends Genet*, 30(6), 211-219. doi: 10.1016/j.tig.2014.04.002
- Tomlins, S. A., Rhodes, D. R., Perner, S., Dhanasekaran, S. M., Mehra, R., Sun, X. W., *et al.* (2005). Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science*, 310(5748), 644-648. doi: 10.1126/science.1117679
- Tonner, E., Barber, M. C., Allan, G. J., Beattie, J., Webster, J., Whitelaw, C. B., *et al.* (2002). Insulin-like growth factor binding protein-5 (IGFBP-5) induces premature cell death in the mammary glands of transgenic mice. *Development*, 129(19), 4547-4557.
- Tonotsuka, N., Hosoi, Y., Miyazaki, S., Miyata, G., Sugawara, K., Mori, T., *et al.* (2006). Heterogeneous expression of DNA-dependent protein kinase in esophageal cancer and normal epithelium. *Int J Mol Med*, 18(3), 441-447.
- Tootle, T. L. and Rebay, I. (2005). Post-translational modifications influence transcription factor activity: a view from the ETS superfamily. *Bioessays*, 27(3), 285-298. doi: 10.1002/bies.20198
- Torchy, M. P., Hamiche, A. and Klaholz, B. P. (2015). Structure and function insights into the NuRD chromatin remodeling complex. *Cell Mol Life Sci*, 72(13), 2491-2507. doi: 10.1007/s00018-015-1880-8
- Toss, A. and Cristofanilli, M. (2015). Molecular characterization and targeted therapeutic approaches in breast cancer. *Breast Cancer Res*, 17, 60. doi: 10.1186/s13058-015-0560-9
- Treilleux, I., Chapot, B., Goddard, S., Pisani, P., Angele, S. and Hall, J. (2007). The molecular causes of low ATM protein expression in breast carcinoma; promoter methylation and levels of the catalytic subunit of DNA-dependent protein kinase. *Histopathology*, 51(1), 63-69. doi: 10.1111/j.1365-2559.2007.02726.x
- Trevino, L. S., Bolt, M. J., Grimm, S. L., Edwards, D. P., Mancini, M. A. and Weigel, N. L. (2016). Differential Regulation of Progesterone Receptor-Mediated Transcription by CDK2 and DNA-PK. *Mol Endocrinol*, 30(2), 158-172. doi: 10.1210/me.2015-1144
- Tropberger, P., Pott, S., Keller, C., Kamieniarz-Gdula, K., Caron, M., Richter, F., *et al.* (2013). Regulation of transcription through acetylation of H3K122 on the lateral surface of the histone octamer. *Cell*, 152(4), 859-872. doi: 10.1016/j.cell.2013.01.032
- Tropberger, P. and Schneider, R. (2013). Scratching the (lateral) surface of chromatin regulation by histone modifications. *Nat Struct Mol Biol*, 20(6), 657-661. doi: 10.1038/nsmb.2581
- Tsukada, Y., Fang, J., Erdjument-Bromage, H., Warren, M. E., Borchers, C. H., Tempst, P., *et al.* (2006). Histone demethylation by a family of JmjC domain-containing proteins. *Nature*, 439(7078), 811-816. doi: 10.1038/nature04433
- Turner, J. D., Vernocchi, S., Schmitz, S. and Muller, C. P. (2014). Role of the 5'-untranslated regions in post-transcriptional regulation of the human glucocorticoid receptor. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, 1839(11), 1051-1061. doi: 10.1016/j.bbagrm.2014.08.010
- Um, J. H., Kang, C. D., Bae, J. H., Shin, G. G., Kim, D. W., Kim, D. W., *et al.* (2004). Association of DNA-dependent protein kinase with hypoxia inducible factor-1 and its implication in resistance to anticancer drugs in hypoxic tumor cells. *Exp Mol Med*, 36, 233-242.
- van Bakel, H. (2011). Interactions of transcription factors with chromatin. In T. R. Hughes (Ed.), *A Handbook of Transcription Factors* Subcellular Biochemistry (Vol. 52, pp. 223-259): Springer. Retrieved from [www.ncbi.nlm.nih.gov/pubmed/21557086](http://www.ncbi.nlm.nih.gov/pubmed/21557086). doi: 10.1007/978-90-481-9069-0\_11
- van den Berg, H. W., Leahey, W. J., Lynch, M., Clarke, R. and Nelson, J. (1987). Recombinant human interferon alpha increases oestrogen receptor expression in human breast cancer cells (ZR-75-1) and sensitizes them to the anti-proliferative effects of tamoxifen. *Br J Cancer*, 55(3), 255-257.
- van der Burg, M., Ijspeert, H., Verkaik, N. S., Turul, T., Wiegant, W. W., Morotomi-Yano, K., *et al.* (2009). A DNA-PKcs mutation in a radiosensitive T-B- SCID patient inhibits Artemis activation and nonhomologous end-joining. *J Clin Invest*, 119(1), 91-98. doi: 10.1172/JCI37141
- van der Knaap, J. A. and Verrijzer, C. P. (2016). Undercover: gene control by metabolites and metabolic enzymes. *Genes Dev*, 30(21), 2345-2369. doi: 10.1101/gad.289140.116
- Vaquerizas, J. M., Kummerfeld, S. K., Teichmann, S. A. and Luscombe, N. M. (2009). A census of human transcription factors: function, expression and evolution. *Nat Rev Genet*, 10(4), 252-263. doi: 10.1038/nrg2538



- Velic, D., Couturier, A. M., Ferreira, M. T., Rodrigue, A., Poirier, G. G., Fleury, F., *et al.* (2015). DNA Damage Signalling and Repair Inhibitors: The Long-Sought-After Achilles' Heel of Cancer. *Biomolecules*, 5(4), 3204-3259. doi: 10.3390/biom5043204
- Venables, J. P., Klinck, R., Koh, C., Gervais-Bird, J., Bramard, A., Inkel, L., *et al.* (2009). Cancer-associated regulation of alternative splicing. *Nat Struct Mol Biol*, 16(6), 670-676. doi: 10.1038/nsmb.1608
- Venkatesan, K., Rual, J.-F., Vazquez, A., Stelzl, U., Lemmens, I., Hirozane-Kishikawa, T., *et al.* (2008). An empirical framework for binary interactome mapping. *Nat Methods*, 6, 83. doi: 10.1038/nmeth.1280
- Venkatesh, S., Smolle, M., Li, H., Gogol, M. M., Saint, M., Kumar, S., *et al.* (2012). Set2 methylation of histone H3 lysine 36 suppresses histone exchange on transcribed genes. *Nature*, 489(7416), 452-455. doi: 10.1038/nature11326
- Venkatesh, S. and Workman, J. L. (2015). Histone exchange, chromatin structure and the regulation of transcription. *Nat Rev Mol Cell Biol*, 16(3), 178-189. doi: 10.1038/nrm3941
- Vermeulen, M., Mulder, K. W., Denissov, S., Pijnappel, W. W., van Schaik, F. M., Varier, R. A., *et al.* (2007). Selective anchoring of TFIID to nucleosomes by trimethylation of histone H3 lysine 4. *Cell*, 131(1), 58-69. doi: 10.1016/j.cell.2007.08.016
- Vettese-Dadey, M., Grant, P. A., Hebbes, T. R., Crane-Robinson, C., Allis, C. D. and Workman, J. L. (1996). Acetylation of histone H4 plays a primary role in enhancing transcription factor binding to nucleosomal DNA in vitro. *EMBO J*, 15(10), 2508-2518.
- Visvader, J. E. (2011). Cells of origin in cancer. *Nature*, 469(7330), 314-322. doi: 10.1038/nature09781
- Visvader, J. E. and Stingl, J. (2014). Mammary stem cells and the differentiation hierarchy: current status and perspectives. *Genes Dev*, 28(11), 1143-1158. doi: 10.1101/gad.242511.114
- Vitale, I., Galluzzi, L., Castedo, M. and Kroemer, G. (2011). Mitotic catastrophe: a mechanism for avoiding genomic instability. *Nat Rev Mol Cell Biol*, 12(6), 385-392. doi: 10.1038/nrm3115
- Vogt, P. K. (2012). Retroviral oncogenes: a historical primer. *Nat Rev Cancer*, 12(9), 639-648.
- Volk, P. and Angrand, P. O. (2007). The control of histone lysine methylation in epigenetic regulation. *Biochimie*, 89(1), 1-20. doi: 10.1016/j.biochi.2006.07.009
- Voss, T. C. and Hager, G. L. (2014). Dynamic regulation of transcriptional states by chromatin and transcription factors. *Nat Rev Genet*, 15(2), 69-81. doi: 10.1038/nrg3623
- Voss, T. C., Schiltz, R. L., Sung, M. H., Yen, P. M., Stamatoyannopoulos, J. A., Biddie, S. C., *et al.* (2011). Dynamic exchange at regulatory elements during chromatin remodeling underlies assisted loading mechanism. *Cell*, 146(4), 544-554. doi: 10.1016/j.cell.2011.07.006
- Wahlström, T. and Arsenian-Henriksson, M. (2015). Impact of MYC in regulation of tumor cell metabolism. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, 1849(5), 563-569. doi: 10.1016/j.bbagr.2014.07.004
- Wang, A., Yue, F., Li, Y., Xie, R., Harper, T., Patel, N. A., *et al.* (2015a). Epigenetic priming of enhancers predicts developmental competence of hESC-derived endodermal lineage intermediates. *Cell Stem Cell*, 16(4), 386-399. doi: 10.1016/j.stem.2015.02.013
- Wang, C., Gong, B., Bushel, P. R., Thierry-Mieg, D., Xu, J., *et al.* (2014). The concordance between RNA-seq and microarray data depends on chemical treatment and transcript abundance. *Nat Biotech*, 32(9), 926-932. doi: 10.1038/nbt.3001
- Wang, C., Mayer, J. A., Mazumdar, A., Fertuck, K., Kim, H., Brown, M., *et al.* (2011). Estrogen Induces c-myc Gene Expression via an Upstream Enhancer Activated by the Estrogen Receptor and the AP-1 Transcription Factor. *Molecular Endocrinology*, 25(9), 1527-1538. doi: 10.1210/me.2011-1037
- Wang, K., Singh, D., Zeng, Z., Coleman, S. J., Huang, Y., Savich, G. L., *et al.* (2010). MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Res*, 38(18), e178. doi: 10.1093/nar/gkq622
- Wang, L., Xie, L., Ramachandran, S., Lee, Y., Yan, Z., Zhou, L., *et al.* (2015b). Non-canonical Bromodomain within DNA-PKcs Promotes DNA Damage Response and Radioresistance through Recognizing an IR-Induced Acetyl-Lysine on H2AX. *Chem Biol*, 22(7), 849-861. doi: 10.1016/j.chembiol.2015.05.014
- Wang, X., Szabo, C., Qian, C., Amadio, P. G., Thibodeau, S. N., Cerhan, J. R., *et al.* (2008). Mutational Analysis of Thirty-two Double-Strand DNA Break Repair Genes in Breast and Pancreatic Cancers. *Cancer Research*, 68(4), 971-975. doi: 10.1158/0008-5472.can-07-6272
- Wang, Z., Gerstein, M. and Snyder, M. (2009a). RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*, 10(1), 57-63.
- Wang, Z., Zang, C., Cui, K., Schones, D. E., Barski, A., Peng, W., *et al.* (2009b). Genome-wide mapping of HATs and HDACs reveals distinct functions in active and inactive genes. *Cell*, 138(5), 1019-1031. doi: 10.1016/j.cell.2009.06.049
- Warnes, G. R., Bolker, B., Bonebakker, L., Gentleman, R., Huber, W., Liaw, A., *et al.* (2015). gplots: Various R Programming Tools for Plotting Data. R package version 2.17.0. Retrieved from <http://CRAN.R-project.org/package=gplots>
- Watanabe, S., Resch, M., Lilyestrom, W., Clark, N., Hansen, J. C., Peterson, C., *et al.* (2010). Structural characterization of H3K56Q nucleosomes and nucleosomal arrays. *Biochim Biophys Acta*, 1799(5-6), 480-486. doi: 10.1016/j.bbagr.2010.01.009
- Watson, L. C., Kuchenbecker, K. M., Schiller, B. J., Gross, J. D., Pufall, M. A. and Yamamoto, K. R. (2013). The glucocorticoid receptor dimer interface allosterically transmits sequence-specific DNA signals. *Nat Struct Mol Biol*, 20(7), 876-883. doi: 10.1038/nsmb.2595
- Watson, P. J., Fairall, L. and Schwabe, J. W. (2012). Nuclear hormone receptor co-repressors: structure and function. *Mol Cell Endocrinol*, 348(2), 440-449. doi: 10.1016/j.mce.2011.08.033
- Wei, G. H., Badis, G., Berger, M. F., Kivioja, T., Palin, K., Enge, M., *et al.* (2010). Genome-wide analysis of ETS-family DNA-binding in vitro and in vivo. *EMBO J*, 29(13), 2147-2160. doi: 10.1038/emboj.2010.106
- Weigel, N. L., Carter, T. H., Schrader, W. T. and O'Malley, B. W. (1992). Chicken progesterone receptor is phosphorylated by a DNA-dependent protein kinase during in vitro transcription assays. *Mol Endocrinol*, 6(1), 8-14. doi: 10.1210/mend.6.1.1738374
- Weigelt, B., Geyer, F. C. and Reis-Filho, J. S. (2010). Histological types of breast cancer: how special are they? *Mol Oncol*, 4(3), 192-208. doi: 10.1016/j.molonc.2010.04.004
- Weikum, E. R., Knuesel, M. T., Ortlund, E. A. and Yamamoto, K. R. (2017). Glucocorticoid receptor control of transcription: precision and plasticity via allostery. *Nat Rev Mol Cell Biol*. doi: 10.1038/nrm.2016.152

- Weirauch, M. T. and Hughes, T. R. (2011). A catalogue of eukaryotic transcription factor types, their evolutionary origin, and species distribution. In T. R. Hughes (Ed.), *A Handbook of Transcription Factors* Subcellular Biochemistry (Vol. 52, pp. 25-73): Springer. Retrieved from [www.ncbi.nlm.nih.gov/pubmed/21557078](http://www.ncbi.nlm.nih.gov/pubmed/21557078). doi: 10.1007/978-90-481-9069-0\_3
- Weterings, E., Verkaik, N. S., Bruggenwirth, H. T., Hoeijmakers, J. H. J. and van Gent, D. C. (2003). The role of DNA dependent protein kinase in synopsis of DNA ends. *Nucleic Acids Research*, 31(24), 7238-7246. doi: 10.1093/nar/gkg889
- Whyte, W. A., Orlando, D. A., Hnisz, D., Abraham, B. J., Lin, C. Y., Kagey, M. H., *et al.* (2013). Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell*, 153(2), 307-319. doi: 10.1016/j.cell.2013.03.035
- Willmore, E., Elliott, S. L., Mainou-Fowler, T., Summerfield, G. P., Jackson, G. H., O'Neill, F., *et al.* (2008). DNA-dependent protein kinase is a therapeutic target and an indicator of poor prognosis in B-cell chronic lymphocytic leukemia. *Clin Cancer Res*, 14(12), 3984-3992. doi: 10.1158/1078-0432.CCR-07-5158
- Wise, D. R. and Thompson, C. B. (2010). Glutamine addiction: a new therapeutic target in cancer. *Trends Biochem Sci*, 35(8), 427-433. doi: 10.1016/j.tibs.2010.05.003
- Wolfrum, C., Besser, D., Luca, E. and Stoffel, M. (2003). Insulin regulates the activity of forkhead transcription factor Hnf-3beta/Foxa-2 by Akt-mediated phosphorylation and nuclear/cytosolic localization. *Proc Natl Acad Sci U S A*, 100(20), 11624-11629. doi: 10.1073/pnas.1931483100
- Wong, R. H., Chang, I., Hudak, C. S., Hyun, S., Kwan, H. Y. and Sul, H. S. (2009). A role of DNA-PK for the metabolic gene regulation in response to insulin. *Cell*, 136(6), 1056-1072. doi: 10.1016/j.cell.2008.12.040
- Woodbine, L., Neal, J. A., Sasi, N. K., Shimada, M., Deem, K., Coleman, H., *et al.* (2013). PRKDC mutations in a SCID patient with profound neurological abnormalities. *J Clin Invest*, 123(7), 2969-2980. doi: 10.1172/JCI67349
- Wu, B., Cao, X., Liang, X., Zhang, X., Zhang, W., Sun, G., *et al.* (2015). Epigenetic regulation of Elf5 is associated with epithelial-mesenchymal transition in urothelial cancer. *PLoS One*, 10(1), e0117510. doi: 10.1371/journal.pone.0117510
- Wu, L., Lee, S. Y., Zhou, B., Nguyen, U. T., Muir, T. W., Tan, S., *et al.* (2013). ASH2L regulates ubiquitylation signaling to MLL: trans-regulation of H3 K4 methylation in higher eukaryotes. *Mol Cell*, 49(6), 1108-1120. doi: 10.1016/j.molcel.2013.01.033
- Wunderlich, Z. and Mirny, L. A. (2009). Different gene regulation strategies revealed by analysis of binding motifs. *Trends Genet*, 25(10), 434-440. doi: 10.1016/j.tig.2009.08.003
- Xie, B. X., Zhang, H., Wang, J., Pang, B., Wu, R. Q., Qian, X. L., *et al.* (2011). Analysis of differentially expressed genes in LNCaP prostate cancer progression model. *J Androl*, 32(2), 170-182. doi: 10.2164/jandrol.109.008748
- Xing, J., Wu, X., Vaporciyan, A. A., Spitz, M. R. and Gu, J. (2008). Prognostic significance of ataxia-telangiectasia mutated, DNA-dependent protein kinase catalytic subunit, and Ku heterodimeric regulatory complex 86-kD subunit expression in patients with nonsmall cell lung cancer. *Cancer*, 112(12), 2756-2764. doi: 10.1002/cncr.23533
- Xu, Q., Li, S., Zhao, Y., Maures, T. J., Yin, P. and Duan, C. (2004). Evidence that IGF binding protein-5 functions as a ligand-independent transcriptional regulator in vascular smooth muscle cells. *Circ Res*, 94(5), E46-54. doi: 10.1161/01.RES.0000124761.62846.DF
- Xu, Y., Zhang, H., Nguyen, Van Thuy M., Angelopoulos, N., Nunes, J., Reid, A., *et al.* (2015). LMTK3 Represses Tumor Suppressor-like Genes through Chromatin Remodeling in Breast Cancer. *Cell Reports*, 12(5), 837-849. doi: 10.1016/j.celrep.2015.06.073
- Yan, C. and Higgins, P. J. (2013). Drugging the undruggable: transcription therapy for cancer. *Biochim Biophys Acta*, 1835(1), 76-85. doi: 10.1016/j.bbcan.2012.11.002
- Yang, C., Shapiro, L. H., Rivera, M., Kumar, A. and Brindle, P. K. (1998). A role for CREB binding protein and p300 transcriptional coactivators in Ets-1 transactivation functions. *Mol Cell Biol*, 18(4), 2218-2229.
- Yang, Y. and Bedford, M. T. (2013). Protein arginine methyltransferases and cancer. *Nat Rev Cancer*, 13(1), 37-50. doi: 10.1038/nrc3409
- Yao, B., Zhao, J., Li, Y., Li, H., Hu, Z., Pan, P., *et al.* (2015). Elf5 inhibits TGF-beta-driven epithelial-mesenchymal transition in prostate cancer by repressing SMAD3 activation. *Prostate*, 75(8), 872-882. doi: 10.1002/pros.22970
- Yao, J. J., Liu, Y., Lacorazza, H. D., Soslow, R. A., Scandura, J. M., Nimer, S. D., *et al.* (2007). Tumor promoting properties of the ETS protein MEF in ovarian cancer. *Oncogene*, 26(27), 4032-4037. doi: 10.1038/sj.onc.1210170
- Yildirim, O., Li, R., Hung, J.-H., Chen, Poshen B., Dong, X., Ee, L.-S., *et al.* (2011). Mbd3/NURD Complex Regulates Expression of 5-Hydroxymethylcytosine Marked Genes in Embryonic Stem Cells. *Cell*, 147(7), 1498-1510. doi: 10.1016/j.cell.2011.11.054
- Yin, J. W. and Wang, G. (2014). The Mediator complex: a master coordinator of transcription and cell lineage development. *Development*, 141(5), 977-987. doi: 10.1242/dev.098392
- You, J. S., Kelly, T. K., De Carvalho, D. D., Taberlay, P. C., Liang, G. and Jones, P. A. (2011). OCT4 establishes and maintains nucleosome-depleted regions that provide additional layers of epigenetic regulation of its target genes. *Proc Natl Acad Sci U S A*, 108(35), 14497-14502. doi: 10.1073/pnas.1111309108
- Young, R. A. (2011). Control of the embryonic stem cell state. *Cell*, 144(6), 940-954. doi: 10.1016/j.cell.2011.01.032
- Yu-Rice, Y., Jin, Y., Han, B., Qu, Y., Johnson, J., Watanabe, T., *et al.* (2016). FOXC1 is involved in ERalpha silencing by counteracting GATA3 binding and is implicated in endocrine resistance. *Oncogene*, 35(41), 5400-5411. doi: 10.1038/onc.2016.78
- Yu, F., Ng, S. S., Chow, B. K., Sze, J., Lu, G., Poon, W. S., *et al.* (2011). Knockdown of interferon-induced transmembrane protein 1 (IFITM1) inhibits proliferation, migration, and invasion of glioma cells. *J Neurooncol*, 103(2), 187-195. doi: 10.1007/s11060-010-0377-4
- Yu, F., Xie, D., Ng, S. S., Lum, C. T., Cai, M. Y., Cheung, W. K., *et al.* (2015). IFITM1 promotes the metastasis of human colorectal cancer via CAV-1. *Cancer Lett*, 368(1), 135-143. doi: 10.1016/j.canlet.2015.07.034
- Yu, J., Vodyanik, M. A., Smuga-Otto, K., Antosiewicz-Bourget, J., Frane, J. L., Tian, S., *et al.* (2007). Induced pluripotent stem cell lines derived from human somatic cells. *Science*, 318(5858), 1917-1920. doi: 10.1126/science.1151526
- Yu, Y., Okayasu, R., Weil, M. M., Silver, A., McCarthy, M., Zabriskie, R., *et al.* (2001). Elevated breast cancer risk in irradiated BALB/c mice associates with unique functional polymorphism of the Prkdc (DNA-dependent protein



- kinase catalytic subunit) gene. *Cancer Res*, 61(5), 1820-1824.
- Yu, Y. L., Chou, R. H., Wu, C. H., Wang, Y. N., Chang, W. J., Tseng, Y. J., *et al.* (2012). Nuclear EGFR suppresses ribonuclease activity of polynucleotide phosphorylase through DNAPK-mediated phosphorylation at serine 776. *J Biol Chem*, 287(37), 31015-31026. doi: 10.1074/jbc.M112.358077
- Zaret, K. S. and Carroll, J. S. (2011). Pioneer transcription factors: establishing competence for gene expression. *Genes Dev*, 25(21), 2227-2241. doi: 10.1101/gad.176826.111
- Zaret, K. S., Lerner, J. and Iwafuchi-Doi, M. (2016). Chromatin Scanning by Dynamic Binding of Pioneer Factors. *Mol Cell*, 62(5), 665-667. doi: 10.1016/j.molcel.2016.05.024
- Zaret, K. S. and Mango, S. E. (2016). Pioneer transcription factors, chromatin dynamics, and cell fate control. *Curr Opin Genet Dev*, 37, 76-81. doi: 10.1016/j.gde.2015.12.003
- Zhang, J., Wu, X. H. and Gan, Y. (2013). Current evidence on the relationship between three polymorphisms in the XRCC7 gene and cancer risk. *Mol Biol Rep*, 40(1), 81-86. doi: 10.1007/s11033-012-2018-9
- Zhang, P., Behre, G., Pan, J., Iwama, A., Wara-Aswapati, N., Radomska, H. S., *et al.* (1999). Negative cross-talk between hematopoietic regulators: GATA proteins repress PU.1. *Proc Natl Acad Sci U S A*, 96(15), 8705-8710.
- Zhang, S., Matsunaga, S., Lin, Y. F., Sishc, B., Shang, Z., Sui, J., *et al.* (2016a). Spontaneous tumor development in bone marrow-rescued DNA-PKcs3A/3A mice due to dysfunction of telomere leading strand deprotection. *Oncogene*, 35(30), 3909-3918. doi: 10.1038/onc.2015.459
- Zhang, T., Cooper, S. and Brockdorff, N. (2015). The interplay of histone modifications - writers that read. *EMBO Rep*, 16(11), 1467-1481. doi: 10.15252/embr.201540945
- Zhang, X., Wang, Y. and Ning, Y. (2017). Down-regulation of protein kinase, DNA-activated, catalytic polypeptide attenuates tumor progression and is an independent prognostic predictor of survival in prostate cancer. *Urologic Oncology: Seminars and Original Investigations*, 35(3), 111.e115-111.e123. doi: 10.1016/j.urolonc.2016.10.012
- Zhang, Y., He, Q., Hu, Z., Feng, Y., Fan, L., Tang, Z., *et al.* (2016b). Long noncoding RNA LINP1 regulates repair of DNA double-strand breaks in triple-negative breast cancer. *Nat Struct Mol Biol*, 23(6), 522-530. doi: 10.1038/nsmb.3211
- Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., *et al.* (2008). Model-based analysis of ChIP-Seq (MACS). *Genome Biol*, 9(9), R137. doi: 10.1186/gb-2008-9-9-r137
- Zhang, Y., Zhang, D., Li, Q., Liang, J., Sun, L., Yi, X., *et al.* (2016c). Nucleation of DNA repair factors by FOXA1 links DNA demethylation to transcriptional pioneering. *Nat Genet*, 48(9), 1003-1013. doi: 10.1038/ng.3635
- Zhang, Z., Wippo, C. J., Wal, M., Ward, E., Korber, P. and Pugh, B. F. (2011). A packing mechanism for nucleosome organization reconstituted across a eukaryotic genome. *Science*, 332(6032), 977-980. doi: 10.1126/science.1200508
- Zhao, B. S., Roundtree, I. A. and He, C. (2017). Post-transcriptional gene regulation by mRNA modifications. *Nat Rev Mol Cell Biol*, 18(1), 31-42. doi: 10.1038/nrm.2016.132
- Zhao, Y., Thomas, H. D., Batey, M. A., Cowell, I. G., Richardson, C. J., Griffin, R. J., *et al.* (2006a). Preclinical evaluation of a potent novel DNA-dependent protein kinase inhibitor NU7441. *Cancer Res*, 66(10), 5354-5362. doi: 10.1158/0008-5472.CAN-05-4275
- Zhao, Y., Yin, P., Bach, L. A. and Duan, C. (2006b). Several Acidic Amino Acids in the N-domain of Insulin-like Growth Factor-binding Protein-5 Are Important for Its Transactivation Activity. *Journal of Biological Chemistry*, 281(20), 14184-14191. doi: 10.1074/jbc.M506941200
- Zhou, J., Ng, A. Y., Tymms, M. J., Jermini, L. S., Seth, A. K., Thomas, R. S., *et al.* (1998). A novel transcription factor, ELF5, belongs to the ELF subfamily of ETS genes and maps to human chromosome 11p13-15, a region subject to LOH and rearrangement in human carcinoma cell lines. *Oncogene*, 17(21), 2719-2732. doi: 10.1038/sj.onc.1202198
- Zhou, Q., Li, T. and Price, D. H. (2012). RNA polymerase II elongation control. *Annu Rev Biochem*, 81, 119-143. doi: 10.1146/annurev-biochem-052610-095910
- Zhou, V. W., Goren, A. and Bernstein, B. E. (2011). Charting histone modifications and the functional organization of mammalian genomes. *Nat Rev Genet*, 12(1), 7-18. doi: 10.1038/nrg2905
- Zhou, W. and Slingerland, J. M. (2014). Links between oestrogen receptor activation and proteolysis: relevance to hormone-regulated cancer therapy. *Nat Rev Cancer*, 14(1), 26-38.
- Zhou, W., Zhu, P., Wang, J., Pascual, G., Ohgi, K. A., Lozach, J., *et al.* (2008). Histone H2A monoubiquitination represses transcription by inhibiting RNA polymerase II transcriptional elongation. *Mol Cell*, 29(1), 69-80. doi: 10.1016/j.molcel.2007.11.002
- Zhou, X., Lindsay, H. and Robinson, M. D. (2014a). Robustly detecting differential expression in RNA sequencing data using observation weights. *Nucleic Acids Res*, 42(11), e91. doi: 10.1093/nar/gku310
- Zhou, Y., Lee, J. H., Jiang, W., Crowe, J. L., Zha, S. and Paull, T. T. (2017). Regulation of the DNA Damage Response by DNA-PKcs Inhibitory Phosphorylation of ATM. *Mol Cell*, 65(1), 91-104. doi: 10.1016/j.molcel.2016.11.004
- Zhou, Z., Patel, M., Ng, N., Hsieh, M. H., Orth, A. P., Walker, J. R., *et al.* (2014b). Identification of synthetic lethality of PRKDC in MYC-dependent human cancers by pooled shRNA screening. *BMC Cancer*, 14, 944. doi: 10.1186/1471-2407-14-944
- Zhu, H., Wang, G. and Qian, J. (2016). Transcription factors as readers and effectors of DNA methylation. *Nat Rev Genet*, 17(9), 551-565. doi: 10.1038/nrg.2016.83
- Zhu, J., Sammons, M. A., Donahue, G., Dou, Z., Vedadi, M., Getlik, M., *et al.* (2015). Gain-of-function p53 mutants co-opt chromatin pathways to drive cancer growth. *Nature*, 525(7568), 206-211. doi: 10.1038/nature15251
- Zumer, K., Low, A. K., Jiang, H., Saksela, K. and Peterlin, B. M. (2012). Unmodified histone H3K4 and DNA-dependent protein kinase recruit autoimmune regulator to target genes. *Mol Cell Biol*, 32(8), 1354-1362. doi: 10.1128/MCB.06359-11