Integrating biospecimen, clinical, imaging and biomarker data to support translational research: A preliminary study in colorectal cancer

Jitendra Jonnagaddala¹, Joanne L Croucher¹,³, Nicola S Meagher¹, Manish Kumar¹, Lena Caruso¹, Toni Rose Jue¹, Vivek Ratnaparkhi⁶
Jack London⁴, Robyn L Ward¹,², Nicholas J Hawkins⁵

¹ Prince of Wales Clinical School, UNSW Australia, 2 Prince of Wales Hospital, Randwick, NSW, Australia, 3 University Library, UNSW Australia, 4 Sidney Kimmel Cancer Center, Thomas Jefferson University, 5 School of Medical Sciences, UNSW Australia, 6 Krishagni Solutions Pty Ltd

Introduction
The Molecular and Cellular Oncology (MCO) cohort represents ~1500 individuals who underwent surgical resection for colorectal cancer between 1994 and 2000 [1]. It contains over 20,000 biospecimens, as well as pathological and treatment data, five years of annual follow up, and a range of molecular analyses including MSI, CIMP, BRAF and KRAS status. To improve accessibility, legacy data had been extracted, curated and loaded into OpenSpecimen (for biospecimen data) or OpenClinica (for clinical data) [2,3]. However, it was clear that integrating these varied data types to extract knowledge would require a more sophisticated translational bioinformatics platform.

Aims
This ANDS-funded project aimed to:
1. Increase access to the valuable MCO cohort data
2. Promote research using the MCO data and specimens within the UNSW Biorepository
3. Demonstrate the value of the tranSMART platform for Australian cancer research

Methods
MCO cohort data from these systems, together with selected TCGA public data [4], was loaded into tranSMART. Developed in 2009, tranSMART is an open source translational bioinformatics platform linked to shared data models and reusable self-service analytics. [5,6]

Use cases were developed relevant to colorectal cancer research and models and mappings for the legacy data were developed with reference to domain standards. Data was then loaded into tranSMART using this bio-curation framework. For privacy reasons, potentially identifiable data were removed or replaced, and some customisations of the tranSMART codebase were implemented.

Results
We have:
- Established the tranSMART bioinformatics platform for the first time in Australia:
- Successfully loaded data from the MCO study and other public datasets into the platform.
- Made de-identified, aggregate-level data publicly accessible via the tranSMART interface at http://tcrn.unsw.edu.au/transmart.
- Made customised features of the tranSMART platform available under open source license. [7]

Interested researchers can use the platform to:
- Define a cohort and locate biospecimens that match their research criteria.
- Generate summary cohort statistics (Figure 1)
- Analyse the cohort data using advanced tools such as survival analysis (Figure 2).
- Compare results across studies using selected TCGA data [4].
- Source the specimens from the UNSW Biorepository, using standard request methods [8].

Next steps
- Integrate additional colorectal cancer and other cancer datasets into the tranSMART platform
- Ensure interested researchers are aware of the MCO data and can access the tranSMART bioinformatics platform

Acknowledgements
This study is supported by the Australian National Data Service (ANDS) through the National Collaborative Research Infrastructure Strategy Program, as well as through the Cancer Institute NSW and UNSW Australia. We also wish to acknowledge the MCO Study Group, the TCGA Research Network and the participants in the MCO Study.

References