

# Rapid detection and identification of foodborne pathogens using genomics

**Author:**

Zhang, Xiaomei

**Publication Date:**

2021

**DOI:**

<https://doi.org/10.26190/unsworks/2056>

**License:**

<https://creativecommons.org/licenses/by/4.0/>

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

Downloaded from <http://hdl.handle.net/1959.4/100146> in <https://unsworks.unsw.edu.au> on 2024-04-16



# **Rapid detection and identification of foodborne pathogens using genomics**

**Xiaomei Zhang**

A thesis in fulfilment of the requirements for  
the degree of Doctor of Philosophy

School of Biotechnology and Biomolecular Sciences  
Faculty of Science

The University of New South Wales

September 2021

## Thesis Title

### Rapid detection and identification of foodborne pathogens using genomics

## Thesis Abstract

Infectious diseases caused by *Salmonella*, *Shigella* and Shiga toxin-producing *E. coli* (STEC) place a heavy burden on human health and incur a massive economic cost. Timely detection and identification of these bacterial pathogens is vital for food safety and public health surveillance. Existing detection methods cannot easily distinguish different serotypes of these pathogens and are time-consuming. Early detection and identification of *Salmonella*, *Shigella* and STEC can be achieved by detection of highly specific and discriminatory pathogen genomic targets. Thus, comparative genomic analysis of many publicly available genomic sequences of *Salmonella*, *Shigella* and STEC has been applied to identify pathogen type-specific gene markers for rapid, highly sensitive and specific identification and differentiation of *Salmonella*, *Shigella* and STEC.

In this thesis, pathogen type-specific gene markers for *Salmonella*, *Shigella* and STEC have been identified through comparative genomic analysis of pathogen genome sequences. For *Salmonella*, a set of 131 serovar-specific genes were identified for prediction of the 106 common serovars from genomic data with 95.3% accuracy. Seven laboratory diagnostic MCDA assays targeting seven *Salmonella* serovar-specific genes were then developed for the detection of five most prevalent *Salmonella* serovars in Australia with high specificity (>93.3%) and high sensitivity (>92.9%). These assays are rapid and can produce results in as short as 8 minutes. For *Shigella*, cluster-specific genes were identified for differentiation of *Shigella* and enteroinvasive *E. coli* (EIEC) from genomic data with 99.64% accuracy and were used to develop an *in silico* pipeline, ShigEiFinder for accurate differentiation, cluster typing and serotyping of *Shigella* and EIEC with 99.38% accuracy. For STEC, cluster/serotype-specific genes were identified for typing of STEC with 99.54% accuracy and were used to develop an *in silico* pipeline, STECFinder which can assign STEC isolates to STEC clusters and serotypes with 99.83% accuracy.

These markers could be adapted for metagenomics or culture independent typing and could also be useful in the development of more cost-effective molecular assays. The outcome of this thesis can be applied to rapid typing of respective pathogens in food, clinical and environmental samples and facilitate surveillance of these pathogens for public health control and prevention.

# THE UNIVERSITY OF NEW SOUTH WALES

## Thesis/Dissertation Sheet

**Surname or Family Name:** Zhang

**Given Name:** Xiaomei

**Other name/s:**

**Abbreviation for degree as give in the University calendar:** PhD

**School:** The School of Biotechnology and Bimolecular Science

**Faculty:** Science

**Thesis Title:** Rapid detection and identification of foodborne pathogens using genomics

### Abstract

Infectious diseases caused by *Salmonella*, *Shigella* and Shiga toxin-producing *E. coli* (STEC) place a heavy burden on human health and incur a massive economic cost. Timely detection and identification of these bacterial pathogens is vital for food safety and public health surveillance. Existing detection methods cannot easily distinguish different serotypes of these pathogens and are time-consuming. Early detection and identification of *Salmonella*, *Shigella* and STEC can be achieved by detection of highly specific and discriminatory pathogen genomic targets. Thus, comparative genomic analysis of many publicly available genomic sequences of *Salmonella*, *Shigella* and STEC has been applied to identify pathogen type-specific gene markers for rapid, highly sensitive and specific identification and differentiation of *Salmonella*, *Shigella* and STEC.

In this thesis, pathogen type-specific gene markers for *Salmonella*, *Shigella* and STEC have been identified through comparative genomic analysis of pathogen genome sequences. For *Salmonella*, a set of 131 serovar-specific genes were identified for prediction of the 106 common serovars from genomic data with 95.3% accuracy. Seven laboratory diagnostic MCDA assays targeting seven *Salmonella* serovar-specific genes were then developed for the detection of five most prevalent *Salmonella* serovars in Australia with high specificity (>93.3%) and high sensitivity (>92.9%). These assays are rapid and can produce results in as short as 8 minutes. For *Shigella*, cluster-specific genes were identified for differentiation of *Shigella* and enteroinvasive *E. coli* (EIEC) from genomic data with 99.64% accuracy and were used to develop an *in silico* pipeline, ShigEiFinder for accurate differentiation, cluster typing and serotyping of *Shigella* and EIEC with 99.38% accuracy. For STEC, cluster/serotype-specific genes were identified for typing of STEC with 99.54% accuracy and were used to develop an *in silico* pipeline, STECFinder which can assign STEC isolates to STEC clusters and serotypes with 99.83% accuracy.

These markers could be adapted for metagenomics or culture independent typing and could also be useful in the development of more cost-effective molecular assays. The outcome of this thesis can be applied to rapid typing of respective pathogens in food, clinical and environmental samples and facilitate surveillance of these pathogens for public health control and prevention.

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

I hereby grant to the University of New South Wales or its agents a non-exclusive licence to archive and to make available (including to members of the public) my thesis or dissertation in whole or in part in the University libraries in all forms of media, now or here after known. I acknowledge that I retain all intellectual property rights which subsist in my thesis or dissertation, such as copyright and patent rights, subject to applicable law. I also retain the right to use all or part of my thesis or dissertation in future works (such as articles or books).

.....  
Signature

.....  
Date

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 can be made when submitting the final copies of your thesis to the UNSW Library. 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

## 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 a non-exclusive licence to archive and to make available (including to members of the public) my thesis or dissertation in whole or part in the University libraries in all forms of media, now or here after known. I acknowledge that I retain all intellectual property rights which subsist in my thesis or dissertation, such as copyright and patent rights, subject to applicable law. I also retain the right to use all or part of my thesis or dissertation in future works (such as articles or books).

For any substantial portions of copyright material used in this thesis, written permission for use has been obtained, or the copyright material is removed from the final public version of the thesis.

## AUTHENTICITY STATEMENT

☒ I certify that the Library deposit digital copy is a direct equivalent of the final officially approved version of my thesis.

# INCLUSION OF PUBLICATIONS STATEMENT

UNSW is supportive of candidates publishing their research results during their candidature as detailed in the UNSW Thesis Examination Procedure.

Publications can be used in the candidate's thesis in lieu of a Chapter provided:

- The candidate contributed **greater than 50%** of the content in the publication and are the "primary author", i.e. they were responsible primarily for the planning, execution and preparation of the work for publication.
- The candidate has obtained approval to include the publication in their thesis in lieu of a Chapter from their Supervisor and Postgraduate Coordinator.
- The publication is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in the thesis.

☒ The candidate has declared that their thesis has publications - either published or submitted for publication - incorporated into it in lieu of a Chapter/s. Details of these publications are provided below.

## Publication Details #1

<b>Full Title:</b>	In silico Identification of Serovar-Specific Genes for Salmonella Serotyping
<b>Authors:</b>	Zhang X, Payne M, Lan R
<b>Journal or Book Name:</b>	Frontier in Microbiology
<b>Volume/Page Numbers:</b>	10:835
<b>Date Accepted/Published:</b>	04.2019
<b>Status:</b>	published

<b>The Candidate's Contribution to the Work:</b>	The candidate is the primary author and contributed > 50% of the content in the publication.
--	--

<b>Location of the work in the thesis and/or how the work is incorporated in the thesis:</b>	This work is incorporated in Chapter 2 in the thesis and presents the first aim of the thesis.
--	--

#### Publication Details #2

<b>Full Title:</b>	Highly Sensitive and Specific Detection and Serotyping of Five Prevalent Salmonella Serovars by Multiple Cross-Displacement Amplification
<b>Authors:</b>	Zhang X, Payne M, Wang Q, Sintchenko V, Lan R
<b>Journal or Book Name:</b>	The Journal of molecular diagnostics
<b>Volume/Page Numbers:</b>	708-19
<b>Date Accepted/Published:</b>	05.2020
<b>Status:</b>	published
<b>The Candidate's Contribution to the Work:</b>	The candidate is the primary author and contributed > 50% of the content in the publication.
<b>Location of the work in the thesis and/or how the work is incorporated in the thesis:</b>	This work is incorporated in Chapter 3 in the thesis and presents the first aim of the thesis.

#### Publication Details #3

<b>Full Title:</b>	Cluster-specific gene markers enhance Shigella and Enteroinvasive Escherichia coli in silico serotyping
<b>Authors:</b>	Zhang X, Payne M, Nguyen T, Kaur S, Lan R
<b>Journal or Book Name:</b>	Microbial genomics
<b>Volume/Page Numbers:</b>	
<b>Date Accepted/Published:</b>	
<b>Status:</b>	submitted
<b>The Candidate's Contribution to the Work:</b>	The candidate is the primary author and contributed > 50% of the content in the publication.
<b>Location of the work in the thesis and/or how the work is incorporated in the thesis:</b>	This work is incorporated in Chapter 4 in the thesis and presents the first aim of the thesis.

#### Publication Details #4

<b>Full Title:</b>	Improved genomic identification, clustering and serotyping of Shiga toxin-producing <i>Escherichia coli</i> using cluster/serotype-specific gene markers
<b>Authors:</b>	Zhang X, Payne M, Kaur S, Lan R
<b>Journal or Book Name:</b>	Frontiers in Cellular and Infection Microbiology
<b>Volume/Page Numbers:</b>	
<b>Date Accepted/Published:</b>	
<b>Status:</b>	submitted
<b>The Candidate's Contribution to the Work:</b>	The candidate is the primary author and contributed > 50% of the content in the publication.
<b>Location of the work in the thesis and/or how the work is incorporated in the thesis:</b>	This work is incorporated in Chapter 5 in the thesis and presents the first aim of the thesis.

#### Candidate's Declaration

I confirm that where I have used a publication in lieu of a chapter, the listed publication(s) above meet(s) the requirements to be included in the thesis. I also declare that I have complied with the Thesis Examination Procedure.



# ACKNOWLEDGEMENTS

The last four years in Lanlab BABS UNSW have been full of unforgettable moments - learning and more learning, warmth, laughter, defeat and struggle. I am finally on my path for a PhD. This significant milestone in my life could not be achieved without the following lovely people, I am grateful to each and every one of you.

First and foremost, I would like to say a very huge thank you to my supervisors Prof. Ruiting Lan and Dr. Michael Payne, since they introduced me to the field of research and inspired my way through this unpredictable PhD journey. I express my deepest, most sincere gratitude to them for their dedicated support and guidance, suggestions for improvement and meticulous proof reading for my thesis. Many thanks to Ruiting for providing this opportunity of PhD study and for his warm encouragement and supervisory. Michael is always supportive and patient, spending tireless time and energy into helping me improve my work. Both of these individuals guided me so positively and always made me feel confident. I learned so much from Ruiting and Michael. Without their guidance and constant support, this PhD would not have been achievable. As one of Ruiting's students and one of Michael's first students, I hope this thesis does their brilliance and patience justice.

I am very grateful to Dr. Laurence Luu who helped me a lot with wet lab work and offered me lots of knowledge. I am profoundly grateful for the hard work of my co-authors to uplift the works presented in this thesis. I am also grateful to the people in Lanlab - Alice Xu, Sandeep Kaur, Ada Luo, Raisa Rafique, Amanda Luo, Thanh Nguyen, Hiroki Suyama and Liam Cheney. Thank you for accompanying me with a friendly and inspiring environment during my PhD journey. I am also very grateful to Liam for proofreading my thesis.

The fourth year of my PhD is without a doubt the hardest. Luckily, I have been surrounded by lovely people in Lanlab who care about each other. Many thanks for your friendship and warm support when I needed most. I would like to give warm thanks and hugs to Alice, Ada and Amanda, for relieving my stress. Thank you for all your support and encouragement during my tough time. Wish you all the best.

At the end of it all, I have the privilege of having a lovely family to support me through the PhD process. My husband, Jack, a heartfelt thank you for being by my side, for always believing in me and encouraging me to follow my dreams. Without your continued love, help, understanding and support with the resources necessary, my PhD would not have been accomplished. Lenny, my precious son, thanks for supporting me with the love and for helping in whatever way. Your love and support are priceless. All of my accomplishments are yours also. Thanks to my family members in China, without their patience and encouragement, I would not have had the courage to embark on this journey.

And finally to Lucky, a lovely, beautiful dog. She has been spending a lot of time with me throughout the past few years, particularly in the pandemic period. Thanks for always accompanying me and bringing me the warmest joys during tough times.

# PUBLICATION AND PRESENTATIONS

## Publications

Zhang X, Payne M, Lan R. *In silico* Identification of Serovar-Specific Genes for *Salmonella* Serotyping. *Front Microbiol.* 2019;10:835.

Zhang X, Payne M, Wang Q, Sintchenko V, Lan R. Highly Sensitive and Specific Detection and Serotyping of Five Prevalent *Salmonella* Serovars by Multiple Cross-Displacement Amplification. *The Journal of molecular diagnostics* : JMD. 2020;22(5):708-19.

Zhang X, Payne M, Nguyen T, Kaur S, Lan R. Cluster-specific gene markers enhance *Shigella* and Enteroinvasive *Escherichia coli* *in silico* serotyping. (Submitted to Microbial genomics 04/02/2021)

Zhang X, Payne M, Kaur S, Lan R. Improved genomic identification, clustering and serotyping of Shiga toxin-producing *Escherichia coli* using cluster/serotype-specific gene markers. (Submitted to Frontiers in Cellular and Infection Microbiology 08/09/2021)

## Presentations

Zhang X, Payne M, Lan R. *In silico* Identification of Serovar-Specific Genes for *Salmonella* Serotyping. (Poster presentation) Australian Society for Microbiology Annual Scientific Meeting 2018.

Zhang X, Payne M, Wang Q, Sintchenko V, Lan R. Highly Sensitive and Specific Detection and Serotyping of Five Prevalent *Salmonella* Serovars by Multiple Cross-Displacement Amplification. (Oral presentation) UNSW BABS symposium 2018.

Zhang X, Payne M, Wang Q, Sintchenko V, Lan R. Highly Sensitive and Specific Detection and Serotyping of Five Prevalent *Salmonella* Serovars by Multiple Cross-Displacement Amplification. (Poster presentation) Australian Society for Microbiology Annual Scientific Meeting 2019.

Zhang X, Payne M, Nguyen T, Kaur S, Lan R. Cluster-specific gene markers enhance *Shigella* and Enteroinvasive *Escherichia coli* *in silico* serotyping. (Poster presentation) Australian Society for Microbiology Annual Scientific Meeting 2021.

# ABSTRACT

Foodborne pathogens can cause foodborne diseases with considerable morbidity and mortality in humans and incur a massive economic cost. A 2010 study found that an estimated 600 million foodborne pathogen infections resulting 420,000 deaths occurred globally. Early detection and identification of contaminating pathogens form a key part of prevention strategy for food safety and public health surveillance. *Salmonella*, *Shigella*, and Shiga toxin-producing *Escherichia coli* (STEC) are the common foodborne bacterial pathogens worldwide. Existing detection and identification methods cannot easily distinguish different serotypes of these pathogens in all cases and are time-consuming. However, timely detection and differentiation of *Salmonella*, *Shigella* and STEC can be achieved by detection of highly specific and discriminatory pathogen genomic targets. With a large number of genomes available, comparative genomic analysis would provide a powerful application to overcome the major challenge in identification of specific genomic targets. Thus, comparative genomic analysis of many publicly available genomic sequences of *Salmonella*, *Shigella* and STEC has been applied in this study for identification of pathogen type-specific gene markers. These gene markers can be used for rapid, highly sensitive and specific identification and differentiation of *Salmonella*, *Shigella* and STEC either from genomic data or using laboratory diagnostic methods.

*Salmonella* is a highly diverse species with more than 2,600 serovars. Only a small proportion of serovars cause severe illness when they contaminate food products. The ability to detect and distinguish this small proportion of illness causing serovars is vital for public health surveillance. In Chapter 2, 106 *Salmonella* serovars covering all of the most common serovars as well as a number of rare serovars were investigated. WGS based phylogenetic analysis showed that there were 81 monophyletic, 24 polyphyletic serovars and one paraphyletic serovar (Enteritidis) among 106 serovars. Comparative genomic analysis of genomic sequences of these *Salmonella* serovars have identified 414 candidate serovar-specific for monophyletic serovars and lineage-specific gene markers for polyphyletic serovars or paraphyletic serovars with 2 or more lineages for these 106 *Salmonella* serovars. This is the largest number of serovar-specific gene markers identified to date. A new approach using the presence or absence of 131 best performing serovar-specific gene markers was designed for molecular *in silico* serotyping of 106

most common *Salmonella* serovars. This approach has an accuracy of 95.3% for *in silico* prediction of the 106 common *Salmonella* serovars from genomic data. The approach can complement current O and H antigen gene based *in silico* serotyping such as SeqSero.

Similar to other parts of the world, *Salmonella* is a common cause of foodborne disease in Australia and over 85% of human *Salmonella* infections in Australia were caused by five *Salmonella* serovars: Typhimurium, Enteritidis, Virchow, Saintpaul, and Infantis. Rapid, accurate and sensitive detection of *Salmonella* and identification of *Salmonella* serovars would be useful for public health investigations. The serovar-specific gene markers with high specificity and sensitivity identified in Chapter 2 were used for the development of more cost-effective laboratory molecular diagnostics assays. The feasibilities of using a cutting edge molecular assay platform to detect these serovar-specific gene markers were conducted in Chapter 3. Seven laboratory diagnostic MCDA assays were developed to detect seven *Salmonella* serovar-specific gene markers for identification of the top five *Salmonella* serovars in Australia. These seven MCDA assays were shown to be highly sensitive (>93.3%) and specific (>93.3%) and can type the five *Salmonella* serovars within 8 minutes. These assays have the potential for culture-independent serotyping of common *Salmonella* serovars directly from clinical samples and showcased the unique applicability of serovar-specific gene markers for rapid detection and serotype identification.

*Shigella* and enteroinvasive *Escherichia coli* (EIEC) cause human bacillary dysentery with similar invasion mechanism. They also share ancestry within *E. coli* with similar physiological, biochemical and genetic characteristics. These similarities make differentiation between *Shigella* and EIEC difficult. However, distinguishing them is important for clinical diagnostic and public health epidemiological investigations. Current genetic markers may not discriminate between *Shigella* and EIEC in all cases. Importantly, multiple phylogenetic clusters identified for *Shigella* and EIEC could provide high resolution separation of *Shigella* and EIEC. In Chapter 4, 10 *Shigella* clusters, 7 EIEC clusters and 53 sporadic types of EIEC were identified by examining over 17,000 publicly available *Shigella* and EIEC genomes. Cluster-specific gene markers for each phylogenetic cluster that was exclusively composed of *Shigella* or EIEC isolates were then identified for differentiation of *Shigella* and EIEC from genomic data

with 99.64% accuracy. A freely available *in silico* serotyping pipeline ShigEiFinder was developed by incorporating the cluster-specific gene markers. ShigEiFinder provided a typing tool for accurate differentiation, cluster typing and serotyping of *Shigella* and EIEC with 99.38% accuracy using genome sequencing data.

STEC infections poses a heavy burden on human health. Detection of STEC infection and determination of the serotype of the causative strain are important for accurate diagnosis and detection of outbreaks for public health control. Current detection and serotyping methods are focused on STEC O157:H7 and “Big 6” non-O157:H7 STEC serotypes. However, other non-O157:H7 STEC serotypes associated with foodborne outbreaks and human infections have been reported frequently in recent years. Therefore, identification of phylogenetic clusters of STEC through large scale examination of publicly available genomes can improve identification and serotyping of STEC by detection of cluster-specific genomic markers. In Chapter 5, 19 STEC major clusters containing O157:H7 and the top 28 non-O157:H7 and 229 STEC minor clusters containing other non-O157:H7 STEC serotypes have been identified through phylogenetic analysis of nearly 41,000 publicly available STEC genomes with 460 different serotypes. Comparative genomic analysis of STEC accessory genomes have identified cluster-specific gene markers for STEC clusters and serotype-specific gene markers for the 10 most common STEC non-O157:H7 for *in silico* typing of STEC with more than 99.54% accuracy. The markers were tested on spiked food metagenomic samples for direct detection and typing of STEC serotypes. Based on these gene markers, an *in silico* pipeline, STECFinder was developed for genomic identification, clustering and serotyping of STEC and has more than 99.65% accuracy.

In conclusion, this thesis has identified highly sensitive and specific pathogen type-specific gene markers for identification and differentiation of serotypes of *Salmonella*, clusters and serotypes of *Shigella*, EIEC and STEC using genomics. These markers could be adapted for metagenomics or culture independent typing and could also be useful in the development of more cost-effective molecular assays. These specific gene markers have been employed to develop genomics based tools for identification of *Salmonella*, *Shigella*, EIEC and STEC clusters and serovars with high specificity and high sensitivity, that can be applied to rapid typing of respective pathogens in food, clinical and

environmental samples and facilitate surveillance of these pathogens for public health control and prevention.

# TABLE OF CONTENTS

<b>ORIGINALITY STATEMENT .....</b>	<b>II</b>
<b>COPYRIGHT STATEMENT .....</b>	Error! Bookmark not defined.
<b>AUTHENTICITY STATEMENT .....</b>	Error! Bookmark not defined.
<b>INCLUSION OF PUBLICATIONS STATEMENT .....</b>	<b>III</b>
<b>ACKNOWLEDGEMENTS.....</b>	<b>VI</b>
<b>PUBLICATION AND PRESENTATIONS.....</b>	<b>VIII</b>
<b>ABSTRACT .....</b>	<b>IX</b>
<b>LIST OF FIGURES .....</b>	<b>XXII</b>
<b>LIST OF TABLES .....</b>	<b>XXIII</b>
<b>LIST OF ABBREVIATIONS .....</b>	<b>XXIV</b>
<b>Chapter 1. Literature review .....</b>	<b>1</b>
1.1 Foodborne pathogens and their burden on human health.....	1
1.2 <i>Salmonella</i> .....	1
1.2.1 Taxonomy, nomenclature and population structure of <i>Salmonella</i> .....	1
1.2.2 Epidemiology .....	4
1.2.2.1 <i>The prevalence of Salmonella subspecies and serovars</i> .....	4
1.2.2.2 <i>The global prevalence of serovars associated with human infections</i> .....	5
1.2.2.3 <i>The prevalence and global distribution in animals and animal-based foods</i> .....	6
1.2.2.4 <i>Transmission modes of Salmonella</i> .....	7
1.2.3 Virulence factors.....	7
1.2.3.1 <i>Protein secretion systems</i> .....	7
1.2.3.2 <i>Toxins</i> .....	8
1.2.3.3 <i>Fimbriae</i> .....	9
1.2.3.4 <i>Flagella</i> .....	9
1.2.4 Mobile genomic elements.....	9
1.2.4.1 <i>Salmonella Pathogenicity Islands (SPIs)</i> .....	9
1.2.4.2 <i>Virulence Plasmids</i> .....	10
1.2.4.3 <i>Prophages</i> .....	13
1.2.4.4 <i>The tRNA<sup>leuX</sup> island</i> .....	14
1.2.4.5 <i>Enteritidis-specific genomic island</i> .....	14



1.2.5 Detection, identification and serotyping of <i>Salmonella</i> .....	15
1.2.5.1 Culture-based methods for detection and phenotypic serotyping.....	15
1.2.5.2 Laboratory methods for molecular serotyping targeting serotype specific <i>O</i> and <i>H</i> antigen genes.....	15
1.2.5.3 Laboratory methods for molecular serotyping targeting genomic markers .....	16
1.2.5.4 Sequence-based molecular subtyping method .....	17
1.2.5.5 WGS based in silico serotyping .....	17
1.2.5.6 Rapid, accurate and sensitive detection of <i>Salmonella</i> using laboratory diagnostic methods.....	18
1.3 <i>Shigella</i> .....	19
1.3.1 Taxonomy and classification .....	19
1.3.1.1 <i>Shigella</i> .....	19
1.3.1.2 Enteroinvasive <i>E. coli</i> (EIEC).....	21
1.3.2 Epidemiology.....	21
1.3.2.1 <i>Shigella</i> .....	21
1.3.2.1.1 Global prevalence .....	21
1.3.2.1.2 High risk population groups.....	22
1.3.2.1.3 Reservoirs and transmission modes .....	22
1.3.2.2 EIEC.....	23
1.3.3 The close relationships between <i>Shigella</i> and EIEC .....	23
1.3.3.1 Phenotypic and biochemical characterization.....	23
1.3.3.2 Genotypic characterization.....	23
1.3.3.3 Virulence of <i>Shigella</i> and EIEC .....	25
1.3.3.4 Phylogenetic relationships.....	25
1.3.3.5 Evolution of <i>Shigella</i> and EIEC .....	26
1.3.4 Detection and identification of <i>Shigella</i> and EIEC.....	27
1.3.4.1 Differentiation of <i>Shigella</i> and EIEC from non-enteroinvasive <i>E. coli</i> ...	27
1.3.4.2 Differentiation of <i>Shigella</i> from EIEC .....	27
1.3.4.3 Differentiation between <i>Shigella</i> species .....	28
1.3.5 Serotyping of <i>Shigella</i> and EIEC.....	29
1.3.5.1 Traditional phenotypic serotyping.....	29
1.3.5.2 Molecular serotyping of <i>Shigella</i> .....	29

1.3.5.3 Molecular serotyping targeting <i>EIEC</i> serotype specific O and H antigen genes.....	30
1.3.5.4 WGS based in silico serotyping .....	30
1.4 STEC .....	31
1.4.1 Classification, nomenclature and population structure of STEC.....	31
1.4.2 Epidemiology.....	32
1.4.2.1 Global incidence .....	32
1.4.2.2 Frequency of diarrhea caused by different serotypes .....	33
1.4.2.3 STEC associated outbreaks.....	34
1.4.2.4 Transmission .....	34
1.4.3 Virulence Factors .....	35
1.4.3.1 Shiga Toxin (Stx).....	35
1.4.3.2 Intimin .....	35
1.4.3.3 Hemolysin.....	36
1.4.4 Mobile genetic elements .....	36
1.4.4.1 Stx Prophages .....	36
1.4.4.2 Locus of enterocyte effacement island .....	37
1.4.4.3 Virulence plasmid O157 (pO157) .....	37
1.4.5 Evolution of STEC and emergence of new STEC serotypes.....	38
1.4.5.1 Overview .....	38
1.4.5.2 Evolution of STEC O157:H7.....	38
1.4.5.3 Hybrid STEC pathotypes.....	38
1.4.6 Detection, identification and serotyping.....	39
1.4.6.1 Conventional culture-based: isolation for typing .....	39
1.4.6.2 Conventional culture-based: confirmation and serotyping .....	40
1.4.6.3 Molecular methods for detection of stx genes and other virulence genes.....	40
1.4.6.4 Molecular serotyping targeting <i>E. coli</i> O and H antigen genes .....	40
1.4.6.5 WGS based in silico serotyping .....	41
1.4.6.6 STEC subtyping in surveillance .....	41
1.5 Metagenomic approaches for detection of foodborne pathogens.....	42
1.6 Limitations of existing methods for detection and serotyping of <i>Salmonella</i> , <i>Shigella</i> and STEC .....	42
1.7 Comparative genomics of accessory genomes .....	43

1.7.1 The accessory genomes of <i>Salmonella</i> , <i>Shigella</i> and STEC .....	43
1.7.2 Comparative genomic analysis of accessory genomes for identification of specific genomic markers .....	43
1.8 Aims of the thesis .....	43
<b>Chapter 2. <i>In silico</i> Identification of Serovar-Specific Genes for <i>Salmonella</i> Serotyping .....</b>	<b>46</b>
2.1 Link to thesis .....	46
2.2 Abstract .....	47
2.3 Introduction .....	48
2.4 Materials and methods.....	49
2.4.1 Ribosomal MLST ST Based Isolate Selection .....	49
2.4.2 Identification of <i>Salmonella</i> Serovar-Specific Candidate Gene Markers.....	50
2.4.3 Evaluation of Potential Serovar-Specific Gene Markers .....	50
2.4.4 Phylogenetic Analyses .....	51
2.4.5 Location and Functions of Serovar-Specific Gene Markers.....	51
2.4.6 <i>In silico</i> Serotype Prediction Using Serovar-Specific Gene Markers.....	51
2.4.7 Calculation of the Specificity of Candidate Serovar-Specific Gene Markers for Common Serovars .....	52
2.5 Results .....	52
2.5.1 Identification of Candidate Serovar-Specific Gene Markers.....	52
2.5.2 Functional Categories of Serovar-Specific Gene Markers .....	54
2.5.3 A Minimal Set of Serovar-Specific Gene Markers for <i>in silico</i> Molecular Serotyping.....	55
2.5.4 Serovar-Specific Gene Markers for Serotyping of Common Serovars .....	59
2.6 Discussion .....	60
2.7 Conclusion.....	65
2.8 Author contributions.....	65
2.9 Funding.....	65
2.10 Supplementary material.....	66
2.11 References .....	66
<b>Chapter 3. Highly sensitive and specific detection and serotyping of five prevalent <i>Salmonella</i> serovars by Multiple Cross Displacement Amplification .....</b>	<b>71</b>
3.1 Link to thesis .....	71

3.2 Abstract .....	72
3.3 Introduction .....	73
3.4 Materials and Methods .....	74
3.4.1 Bacterial strains and Genomic DNA extraction .....	74
3.4.2 Design of MCDA primers and the specificities of MCDA products.....	74
3.4.3 The initial evaluation of the seven MCDA assays.....	79
3.4.4 Evaluation of the limit of detection of the MCDA assays in pure culture.....	79
3.4.5 Evaluation of the sensitivity and specificity of seven MCDA assays in pure culture .....	80
3.4.6 Phylogenetic analyses .....	80
3.5 Results .....	81
3.5.1 Selection of serovar specific genes for MCDA products targeting five serovars .....	81
3.5.2 Evaluation of limit of detection (LOD) of the seven MCDA assays in pure culture .....	82
3.5.3 Evaluation of the sensitivity of seven MCDA assays in pure culture .....	82
3.5.4 Evaluation of the specificity of seven MCDA assays in pure culture .....	86
3.6 Discussion .....	86
3.7 Acknowledgements .....	91
3.8 Supplemental Data .....	92
3.9 References .....	92
<b>Chapter 4. Cluster-specific gene markers enhance <i>Shigella</i> and Enteroinvasive <i>Escherichia coli</i> in silico serotyping.....</b>	<b>99</b>
4.1 Link to Thesis.....	99
4.2 Abstract .....	100
4.3 Introduction .....	100
4.4 Materials and Methods .....	103
4.4.1 Identification of <i>Shigella</i> and EIEC isolates from NCBI database.....	103
4.4.2 Genome sequencing.....	103
4.4.3 Genome assembly and data processing .....	104
4.4.4 Selection of isolates for <i>Shigella</i> and EIEC identification dataset .....	104
4.4.5 Phylogeny of <i>Shigella</i> and EIEC based on WGS .....	105
4.4.6 Investigation of <i>Shigella</i> virulence plasmid pINV .....	105

4.4.7 Identification of the cluster-specific gene markers.....	105
4.4.8 Validation of the cluster-specific gene markers .....	106
4.4.9 Development of ShigEiFinder, an automated pipeline for molecular serotyping of <i>Shigella</i> and EIEC .....	106
4.5 Results .....	107
4.5.1 Screening sequenced genomes for <i>Shigella</i> and EIEC isolates .....	107
4.5.2 Identification of <i>Shigella</i> and EIEC clusters.....	108
4.5.3 Analysis of the 59 sporadic EIEC isolates.....	113
4.5.4 Identification of cluster-specific gene markers.....	114
4.5.5 Validation of cluster-specific gene markers .....	116
4.5.6 Development of an automated pipeline for molecular serotyping of <i>Shigella</i> and EIEC.....	118
4.5.7 The accuracy and specificity of ShigEiFinder in cluster typing.....	119
4.5.8 Comparison of ShigEiFinder and ShigaTyper.....	122
4.6 Discussion .....	123
4.6.1 Determining phylogenetic clusters for better separation of <i>Shigella</i> isolates from EIEC.....	123
4.6.2 Highly sensitive and specific cluster-specific gene markers for differentiation of <i>Shigella</i> and EIEC isolates .....	124
4.6.3 ShigEiFinder can accurately type <i>Shigella</i> and EIEC .....	125
4.7 Conclusion.....	126
4.8 Authors and contributors .....	126
4.9 Acknowledgements .....	126
4.10 Data bibliography .....	126
4.11 Abbreviations .....	127
4.12 References .....	127
<b>Chapter 5. Improved genomic identification, clustering and serotyping of Shiga toxin-producing <i>Escherichia coli</i> using cluster/serotype-specific gene markers...</b>	<b>136</b>
5.1 Link to thesis .....	136
5.2 Abstract .....	136
5.3 Introduction .....	137
5.4 Materials and Methods .....	140
5.4.1 Identification of STEC isolates from NCBI database.....	140

5.4.2 Genome assembly and data processing .....	140
5.4.3 Selection of isolates for STEC identification dataset .....	141
5.4.4 Phylogeny of STEC isolates based on WGS .....	141
5.4.5 Identification of the cluster/serotype-specific gene markers .....	142
5.4.6 Validation of the cluster/serotype-specific gene markers.....	142
5.4.7 Detection of the cluster/serotype-specific gene markers in STEC spiked food samples using shotgun metagenomic sequencing reads .....	143
5.4.8 Development of STECFinder, an automated pipeline for molecular serotyping of STEC .....	143
5.5 Results .....	145
5.5.1 Screening sequenced genomes for STEC isolates .....	145
5.5.2 The frequency of STEC serotypes .....	145
5.5.3 Identification of STEC clusters.....	146
5.5.4 Identification of the cluster/serotype-specific gene markers .....	150
5.5.5 Validation of cluster/serotype-specific gene markers.....	152
5.5.6 Detection of the cluster/serotype-specific gene markers in the spiked food samples using shotgun metagenomic sequencing reads .....	153
5.6 Discussion .....	157
5.7 Conclusion.....	160
5.8 Conflict of Interest.....	160
5.9 Author Contributions.....	160
5.10 Abbreviations .....	161
5.11 Supplementary tables: .....	161
5.12 References .....	161
<b>Chapter 6. General Discussion.....</b>	<b>173</b>
6.1 Key findings and significance of this study .....	173
6.1.1 Key findings of this study .....	173
6.1.1.1 <i>Salmonella</i> serovar-specific gene markers identified for most frequent <i>Salmonella</i> serovars .....	173
6.1.1.2 <i>Seven MCDA assays developed for highly sensitive and specific detection and serotyping of five prevalent Salmonella serovars .....</i>	173
6.1.1.3 <i>Cluster-specific gene markers identified for differentiation of Shigella and EIEC.....</i>	174

6.1.1.4 Cluster/serotype-specific gene markers identified for identification, clustering and serotyping of STEC .....	175
6.1.2 Significance of this study .....	175
6.2 Establishment of high quality and representative WGS data for identification of pathogen type-specific gene markers .....	176
6.3 Establishing a systematic approach for identification of pathogen type-specific gene markers.....	177
6.3.1 Establishment of a systematic approach .....	177
6.3.2 Pathogen type-specific gene marker sets increase the sensitivity and specificity of typing .....	178
6.4 WGS based phylogenetic analysis for <i>Salmonella</i> , <i>Shigella</i> and STEC .....	179
6.4.1 <i>Salmonella</i> serovar diversity.....	179
6.4.2 WGS based analysis identified phylogenetic clusters of <i>Shigella</i> and EIEC.....	179
6.4.3 WGS based analysis identified phylogenetic clusters of STEC .....	180
6.5 <i>Salmonella</i> serovar prediction using serovar-specific gene markers can enhance or replace existing molecular serotyping methods .....	181
6.6 <i>Salmonella</i> serovar-specific gene markers can be used to predict major serovars across the globe .....	181
6.7 <i>Salmonella</i> serovar-specific gene markers can be used to develop laboratory detection and serotyping assays .....	182
6.8 Cluster-specific gene markers enhance <i>Shigella</i> and EIEC differentiation and serotyping .....	184
6.9 Cluster/serotype-specific gene markers improve STEC identification, clustering and serotyping .....	185
6.10 STEC cluster/serotype-specific gene markers can be adapted for metagenomics based diagnosis for rapid STEC identification.....	186
6.11 Future directions and serotyping of <i>Salmonella</i> , <i>Shigella</i> and STEC .....	187
6.12 Conclusion.....	188
REFERENCES.....	189
APPENDIX .....	243
Appendix I: Supplementary Material of Chapter 2 .....	243
Appendix II: Supplementary Material of Chapter 3.....	244
Appendix II: Table S1: Bacterial strains used in this study.....	244

Appendix II: Table S2: <i>in silico</i> sensitivity and specificity of the seven MCDA products.....	247
Appendix II: Data S1: The sequences of seven serovar/lineage-specific gene markers.....	248
Appendix II: Figure S1: Phylogenetic relationship of 4 SARB Enteritidis strains. ....	251
Appendix II: Figure S2: Phylogenetic relationship of SARB17, Enteritidis and other serovars. ....	252
Appendix II: Figure S3: Phylogenetic relationship of Infantis strain SARB27. ...	253
Appendix III: Supplementary Material of Chapter 4 .....	254
Appendix III: Data S1 Additional scripts information .....	254
Appendix III: Data S2 Algorithms incorporated into the ShigEiFinder.....	256
Appendix III: Data S3 <i>Shigella</i> /EIEC serotypes specific O and H antigens used in ShigEiFinder .....	265
Appendix III: Figure S1: Identification phylogenetic tree.....	270
Appendix III: Figure S2-A: Confirmation phylogenetic tree .....	271
Appendix III: Figure S2-B: Confirmation phylogenetic tree.....	272
Appendix III: Figure S3: Distribution of mapped 38 virulence genes in 58 sporadic isolates. ....	273
Appendix III: Figure S4: Validation phylogenetic trees.....	274
Appendix IV: Supplementary Material of Chapter 5 .....	281
Appendix IV: Supplementary Tables: .....	281
Appendix IV: Figure S1: Identification phylogenetic tree .....	282
Appendix IV: Figure S2: Validation phylogenetic tree .....	283



# LIST OF FIGURES

Figure 1.2-1: <i>Salmonella</i> genus nomenclature. Adapted from Ryan et al. [14].....	2
Figure 1.3-2: The pINV plasmid of <i>Shigella</i> and EIEC. Adopted from Pasqua et al. [242]. .....	24
Figure 1.4-1: Pathotypes of <i>E. coli</i> .....	32
Chapter 2. Figure 1: The distribution of sensitivity and specificity of 354 potential serovar-specific gene markers.....	53
Chapter 2. Figure 2: The distribution of a minimal set of 131 serovar-specific genes in 106 serovars. ....	55
Chapter 3. Figure 1: Nucleotide sequence and location of seven Multiple Cross Displacement Amplification (MCDA) primers' sets. ....	75
Chapter 3. Figure 2: Limit of Detection (LoD) amplification curves of seven Multiple Cross Displacement Amplification (MCDA) assays. ....	83
Chapter 3. Figure 3: Standard curves of seven Multiple Cross Displacement Amplification (MCDA) assays based on average detection times and serial dilutions. ....	84
Chapter 4. Figure 1: <i>Shigella</i> and EIEC cluster identification phylogenetic tree. ....	110
Chapter 4. Figure 2: <i>in silico</i> serotyping pipeline workflow. ....	117
Chapter 5. Figure 1: The frequency of 463 STEC serotypes. ....	146
Chapter 5. Figure 2: STEC cluster identification phylogenetic tree. ....	149
Chapter 5. Figure 3: The frequency of STEC serotypes (O157:H7 and top 18 non- O157:H7) in STEC clusters. ....	151
Chapter 5. Figure 4: <i>in silico</i> serotyping pipeline workflow.. ....	155

# LIST OF TABLES

Table 1.2-1: Summary of <i>Salmonella</i> pathogenicity islands (SPIs) .....	11
Chapter 2. Table 1: Lineage-specific candidate gene markers for polyphyletic serovars and paraphyletic serovar .....	56
Chapter 2. Table 2: Serovar-specific genes functional categories .....	58
Chapter 2. Table 3: A panel of serovar-specific genes for typing the ten most frequent serovars in Australia.....	61
Chapter 3. Table 1: Primers used for seven Multiple Cross Displacement Amplification assays .....	76
Chapter 3. Table 2: The sensitivity and specificity (%) of the seven MCDA assays .....	89
Chapter 4. Table 1: The summary of identified <i>Shigella</i> and EIEC clusters and outliers in identification dataset .....	112
Chapter 4. Table 2: The sensitivity and specificity of cluster-specific genes.....	115
Chapter 4. Table 3: The accuracy of ShigEiFinder with identification dataset and validation dataset.....	120
Chapter 4. Table 4: The assignments of 15,501 validation isolates by ShigEiFinder and Shigatyper .....	121
Chapter 5. Table 1: Major STEC clusters identified in identification dataset .....	148
Chapter 5. Table 2: Summary of identified STEC minor clusters in identification dataset .....	151
Chapter 5. Table 3: The sensitivity and specificity of STEC cluster/serotype-specific gene markers .....	154

# LIST OF ABBREVIATIONS

A/E lesions	Attaching and effacing lesions
CDT	Cytolethal distending toxin
CdtB	Cytolethal distending toxin subunit B
cgMLST	core genome MLST
CIT	Culture-independent testing
CS54	Genetic island located at centisome 54
DALYs	Disability Adjusted Life Years
<i>eae</i>	intimin
EAEC	Enteraggregative <i>E. coli</i>
ECOR	<i>Escherichia coli</i> reference collection
EHEC	Enterohemorrhagic <i>E. coli</i>
EHEC-Hly	EHEC hemolysin
EIEC	Enteroinvasive <i>E. coli</i>
ELISA	Enzyme-linked immunosorbent assay
ENA	European Nucleotide Archive
EPEC	Enteropathogenic <i>E. coli</i>
ESRD	End-stage renal disease
ETEC	Enterotoxigenic <i>E. coli</i>
ExPEC	Extraintestinal pathogenic <i>E. coli</i>
FD	Fluorescent dye
FN	False negatives
FP	False positives
FPR	False positive rate
GEI	Enteritidis-specific genomic island
HC	Haemorrhagic colitis
HK	House Keeping
HUS	Haemolytic uraemic syndrome

iNTS	invasive NTS
<i>ipaH</i>	Invasion plasmid antigen H
LAMP	Loop-mediated isothermal amplification
<i>lacY</i>	Lactose permease
LDC	Lysine-decarboxylase
LEE	Locus of enterocyte effacement
LoD	Limit of detection
LPS	Lipopolysaccharide
MAC	Major Antigenic Cluster
MCDA	Multiple cross displacement amplification
MGE	Mobile genomic elements
MLST	Multilocus sequence typing
MLVA	Locus variable-number tandem repeat analysis
MPI	Major pathogenicity island
MSM	Men who have sex with men
NCBI	National Center for Biotechnology Information
NEPSS	National Enteric Pathogens Surveillance Scheme
NTS	Non-typhoidal <i>Salmonella</i>
<i>oac</i>	O-acetylation
OMV	Outer membrane vesicles
<i>opt</i>	Phosphoethanolamine transferase
ORF	Open reading frame
PCR	Polymerase chain reaction
<i>pef</i>	Plasmid-encoded fimbriae
PFGE	Pulsed-field gel electrophoresis
pINV	<i>Shigella</i> virulence plasmid
PPV	Positive predictive value
pSLT	<i>Salmonella</i> Virulence Plasmid
<i>rfb</i>	O antigen gene cluster

rfb-RFLP	Restriction fragment length polymorphism
rMLST	ribosomal MLST
rSTs	ribosomal MLST STs
RTX	Repeats-in-toxin
SARA	<i>Salmonella</i> reference collections A
SARB	<i>Salmonella</i> reference collections B
SARC	<i>Salmonella</i> reference collections C
SB	<i>Shigella boydii</i>
SD	<i>Shigella dysenteriae</i>
SD1	<i>Shigella dysenteriae</i> serotype 1
Sdf I	<i>Salmonella</i> difference fragment
SF	<i>Shigella flexneri</i>
SHI PAI	<i>Shigella</i> -specific pathogenicity islands
ShigEiFinder	<i>Shigella</i> EIEC Cluster Enhanced Serotype Finder
SISTR	<i>Salmonella in silico</i> Typing Resource
SMAC	Sorbitol-containing MacConkey agar
SNP	Single nucleotide polymorphism
SPIs	<i>Salmonella</i> pathogenicity islands
SRA	Sequence Read Archive
SS	<i>Shigella sonnei</i>
ST	Sequence types
STEC	Shiga toxin-producing <i>E. coli</i>
Stn	<i>Salmonella</i> enterotoxin
STV	<i>Salmonella</i> Typing Virulence
Stx	Shiga Toxin
T1SS	Type I secretion system
T3SS	Type III secretion system
T6SS	Type VI secretion system
TN	True negatives

TNR	True negative rate
TP	True positives
TPR	True positive rate
TSI	Triple sugar iron agar
<i>uidA</i>	$\beta$ -glucuronidase
wgMLST	Whole genome MLST
WGS	Whole-genome sequencing
<i>wzx</i>	O-antigen flipase gene
<i>wzy</i>	O-antigen polymerase gene

# Chapter 1. Literature review

## 1.1 Foodborne pathogens and their burden on human health

Foodborne pathogens are microbiological agents (e.g. viruses, bacteria, parasites) that can cause foodborne diseases [1]. Food products contaminated with foodborne pathogens result in considerable morbidity and mortality in humans [2]. Foodborne diseases present a major public health problem worldwide, particularly in children under 5 years old [2,3]. Globally, an estimated 582 million foodborne pathogen infections resulted in 25.2 million Disability Adjusted Life Years (DALYs) and 351,000 deaths in 2010 [2]. Foodborne diarrheal disease agents were the most frequent cause of foodborne diseases and caused 550 million foodborne pathogen infections, 15.8 million DALYs and 200,000 deaths in 2010 [2]. Of the foodborne diarrheal disease agents, bacterial are the major cause of foodborne diseases, in particular *Salmonella*, *Shigella* and Shiga toxin-producing *E. coli* (STEC) [2,3] are common.

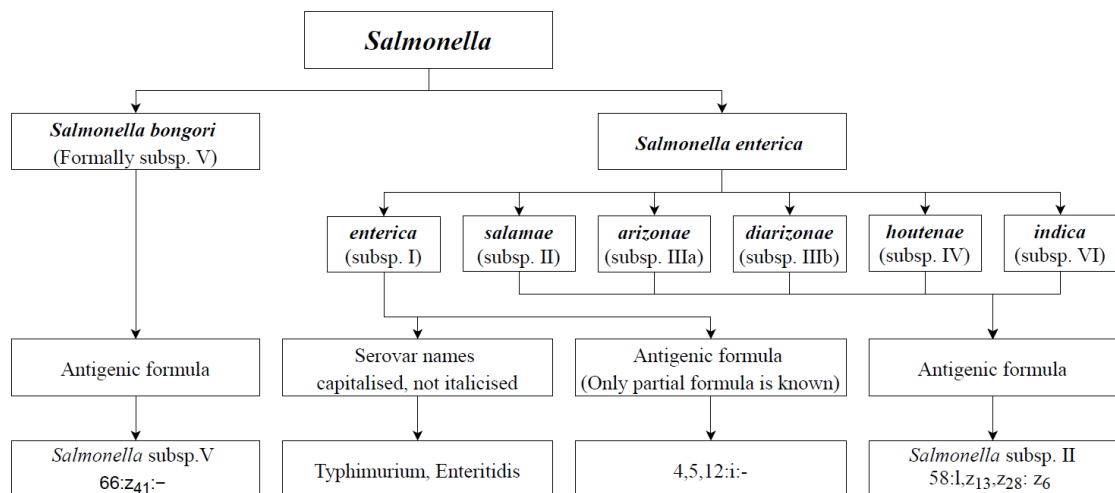
## 1.2 *Salmonella*

*Salmonella* is gram negative, rod-shaped, non-spore-forming, facultative anaerobe, oxidase negative and mobile by peritrichous flagella [4]. *Salmonella* is responsible for the second most common gastrointestinal human infections caused by foodborne bacterial pathogens [2]. *Salmonella* causes human salmonellosis characterised by enteric fever and diarrheal [5] and is responsible for the most foodborne DALYs [3,6].

### 1.2.1 Taxonomy, nomenclature and population structure of *Salmonella*

*Salmonella* belongs to the Enterobacteriaceae family and is divided into two species, *Salmonella enterica* and *Salmonella bongori* (formally classified as *S. enterica* subspecies V) [7]. *S. enterica* is further divided into 6 subspecies: *S. enterica* subsp. *enterica* (I), *S. enterica* subsp. *salamae* (II), *S. enterica* subsp. *arizonae* (IIIa), *S. enterica* subsp. *diarizonae* (IIIb), *S. enterica* subsp. *houtenae* (IV), and *S. enterica* subsp. *indica* (VI) based on biochemical and genomic characteristics [8,9]. *Salmonella* is further classified into more than 2,600 serotypes (also known as serovars) according to antigenic classification system used in White-Kauffmann- Le Minor Scheme [10-12]. Among which, over 1,500 serovars belong to *S. enterica* subsp. *enterica* (I) [12] (Figure 1.2-1)

*Salmonella* has somatic O antigen, flagellar H antigen and Vi capsular antigen [5,13,14]. Somatic O antigen is the variable polysaccharide in the outer surface of the lipopolysaccharide (LPS) and consists of oligosaccharide repeats (O units) which are responsible for O antigen specificity [5,15]. O antigen biosynthesis proteins are encoded by the *wzx* (O-antigen flipase gene) and the *wzy* (O-antigen polymerase gene) genes located on the O antigen gene cluster (*rfb*) which are highly specific for the majority of O groups [15,16]. Flagellar H antigen has two forms named phase 1 (H1) and phase 2 (H2). H1 and H2 antigens are encoded by *fliC* and *fljB* genes, respectively [17]. There are 46 O antigens and 119 H antigens described in the Kauffmann–White–Le Minor serotyping scheme [12]. Vi capsular antigen is a polysaccharide encoded by *viaA* and *viaB* on the chromosome and is present in 3 serovars only [5,18].



**Figure 1.2-1: *Salmonella* genus nomenclature.** Adapted from Ryan *et al.* [14]

Serovar is designated based on the combination of O and H antigens described in the Kauffmann–White–Le Minor serotyping scheme [10]. The unique combination of O, H1 and H2 antigens provides the antigenic formular which is referring to a serovar name. Each antigen in an antigenic formula is separated by a colon (O:H1:H2). For example, in the antigenic formular 4,5,12:i:-, “4,5,12” is O antigen factors, “i” is H1 antigen and H2 antigen is absent. An antigenic formula is assigned to all serovars according to the Kauffmann-White-Le Minor scheme [12]. Serovars expressing both H1 and H2 antigens are called diphasic, while serovars expressing only one type of H antigen is called



monophasic [19]. Some strains may lose O or H antigen expression resulting in rough or nonmotile strains respectively [20].

Common names are also assigned to many serovars in *S. enterica* subsp. *enterica* and often refer to the geographic location where the serovar was first isolated or describe an aspect of the serovars pathogenicity [8]. The first letter of serovar name is capitalized and the full name is not italicized. The full name “*Salmonella enterica* subsp./ssp. *enterica* serovar followed by name of serovar” is used in the first mention in text, for example, *Salmonella enterica* subsp./ssp. *enterica* serovar Enteritidis. Subsequently, the name can be written with the genus ‘*Salmonella*’ followed by the name of serovar (*Salmonella* Enteritidis) [8,14]. In contrast, the antigenic formulars are assigned to serovars for only partial formula of *S. enterica* subsp. *enterica* and the other five *S. enterica* subspecies and *S. bongori* [8,14]. The full name can be written in subspecies (Roman letters, not italicised) along with antigenic formular, for example, *Salmonella* subsp. II 58: 1,z<sub>13</sub>,z<sub>28</sub>: z<sub>6</sub> [14] (Figure 1.2-1).

In addition to the White-Kauffmann- Le Minor Scheme, the Major Antigenic Cluster (MAC) types have been used in *Salmonella* nomenclature [21]. MAC types are defined by the sequence types (ST) assigned by the Multilocus sequence typing (MLST) [22] and genetic antigenic profile of serovars in the White-Kauffmann- Le Minor Scheme [23]. *Salmonella* serovars are named by *Salmonella* species, subspecies and then MAC type [21]. For example, *Salmonella* Typhimurium in MAC type is written with *S. enterica* ST34—*S. Typhimurium*.

There are three *Salmonella* reference collections A, B and C (SARA, SARB and SARC) that have been established for use for research purposes [24-26]. SARA is a reference collection of 72 strains of the *Salmonella* Typhimurium complex representing the serovars Typhimurium, Saintpaul, Heidelberg, Paratyphi B (including variety java) and Muenchen [24]. SARB consists of 72 phylogenetically well-characterized strains belonging to 37 serovars of *Salmonella* subspecies *enterica* [25]. Lastly, SARC is a reference collection of 16 strains of *S. enterica* (all six subspecies) and *S. bongori* [26].

From a phylogenetic perspective, *Salmonella enterica* subsp. *enterica* serovars are classified into monophyletic, polyphyletic and paraphyletic serovars. All members of a monophyletic serovar are found within a single clade that only contains that serovar with a single common ancestor. A polyphyletic serovar contains members with different common ancestors which are separated by clades of other serovars. A paraphyletic serovar contains a common ancestor like a monophyletic serovar but a subset of the clade is a different serovar. A recent phylogenetic analysis reported that ~10% of 266 different serovars investigated are polyphyletic or paraphyletic [27].

## **1.2.2 Epidemiology**

### ***1.2.2.1 The prevalence of Salmonella subspecies and serovars***

*S. bongori* and all *S. enterica* subspecies can cause salmonellosis in humans and animals. *S. enterica* subsp. *enterica* is the cause of over 99% of human and warm-blooded animals' salmonellosis and only a small proportion of serovars cause human salmonellosis [6,28-30]. *S. bongori* and the other five *S. enterica* subspecies are often related to infections in cold-blooded animals such as reptiles and snakes [14].

On the basis of host specificity, *S. enterica* subsp. *enterica* serovars can be classified into human-restricted serovars (Typhi, Sendai, Paratyphi A, Paratyphi B and Paratyphi C), animal-adapted serovars (Dublin in cattle, Gallinarum/ Pullorum in poultry, Abortusovis in sheep, Choleraesuis in pigs and Abortusequi in horses) and broad-host serovars (such as Typhimurium in human, cattle and pigs, Enteritidis in human and chicken) [23,31,32].

In human salmonellosis, *S. enterica* subsp. *enterica* serovars are divided into typhoidal serovars (Typhi, Paratyphi A, Paratyphi B, Paratyphi C) and non-typhoidal *Salmonella* (NTS) serovars. Typhoidal serovars cause enteric fevers (typhoid and paratyphoid fevers) and invasive diseases, while NTS serovars cause diarrheal diseases and invasive diseases [31]. The NTS serovars causing invasive diseases are referred as invasive NTS (iNTS) serovars. Among NTS serovars, only a small proportion are responsible for human salmonellosis [31,33].

### ***1.2.2.2 The global prevalence of serovars associated with human infections***

*S. enterica* is the second leading cause of bacterial foodborne diseases (including both sporadic and outbreak cases) worldwide [3]. Globally, an estimated 93.8 million NTS infections with 155,000 deaths occur each year, of these, 80.3 million are considered foodborne [34]. Annually, the estimated cases of iNTS infections were 3.4 million, resulting in 681,000 deaths worldwide [35]. The global incidence of NTS and iNTS infections were 1,140 cases, and 4 cases per 100,000 people respectively [2,35]. Meanwhile, an estimated foodborne typhoidal fever occurred in 2010 was 9.3 million cases, leading to 64,000 deaths and 4.6 million DALYs. The global incidence of foodborne typhoidal fever was 135 cases per 100,000 people [2].

*Salmonella* infection types differ across geographic regions and populations. NTS infections occur globally, whereas iNTS infections are more prevalent in sub-Saharan Africa and Southeast Asia [31,36-39]. The majority of NTS infections occur in individuals over 5 years old and most iNTS infections are associated with individuals under 5 years old and over 65 years old [3]. Typhoidal fevers are prevalent in developing countries particularly in Africa [2,31,38]. Typhoidal fever is common among children under 12 years old in regions with high-incidence and occurs in all age groups in low-incidence regions [40,41]. Human infections caused by *S. bongori* and the other five *S. enterica* subspecies are very uncommon and infections mainly affect children aged 1 month to 3 years [28,29].

In Australia, foodborne human *Salmonella* infections was estimated at 185 per 100,000 population each year with a proportion of cases linked to outbreaks [42,43]. Typhimurium was the most prevalent serovar and was responsible for 43.9% of human salmonellosis and 84% of foodborne *Salmonella* outbreaks between 2001 and 2016 [44]. Enteritidis is the second most prevalent serovar. In this case however, human infections are mostly acquired overseas [43]. The Virchow and Saintpaul serovars are ranked third and fourth most prevalent but are less common in other countries. Infantis is the fifth most frequently reported serovar [43,45].

In the United States, *Salmonella* causes the majority of bacterial foodborne diseases [46,47]. The incidence of *Salmonella* infections is 1,002 cases per 100,000 population

each year [2]. The annual cases were estimated to be 1.2 million with 452 deaths, of which, up to 70% of cases were caused by the top 20 NTS serovars [47] and 50% of all cases were caused by the top 6 serovars [45]. The most common serovars causing human infection were Enteritidis, followed by Typhimurium, Newport, Javiana and Typhimurium monophasic variants 1,4,[5],12:i: according to annual culture-confirmed surveys [46].

In Europe, an estimated 5.1 million foodborne *Salmonella* infections occur annually [2,47]. Enteritidis, Typhimurium and Typhimurium monophasic variants 1,4,[5],12:i:- were the three most frequently reported serovars, accounting for 50.3%, 11.9% and 8.2% of human cases respectively. Infantis and Newport were the fourth and fifth most commonly reported serovars [48,49].

#### ***1.2.2.3 The prevalence and global distribution in animals and animal-based foods***

Animal-adapted serovars and broad-host serovars cause infections in farm animals. The prevalence of *Salmonella* in farm environments ranges from 10% to 26% [50].

In pigs and pork, the prevalence of *Salmonella* varies from 3% to 33% [4,51,52]. Regarding serovars associated with pigs, Typhimurium and Derby are frequently reported in Europe, Oceania, Asia, and North America. Sofia and Kentucky are frequently reported in Oceania and North America, respectively [4].

In poultry, the prevalence of *Salmonella* ranges between 5% and 100% among various environmental and faecal samples [4]. In chicken associated serovars, Enteritidis is frequently reported in Asia, Africa, United States, Europe and Latin America; Sofia is frequently reported in Oceania [51].

In cattle and beef, the prevalence of *Salmonella* is 8.5% [50]. In cattle and beef associated serovars, Anatum and Typhimurium are most frequently reported in Africa, Latin America, and Europe. Agona and Muenchen are frequently reported in Asia and Oceania, respectively [51].

In seafood, the prevalence of *Salmonella* is 12% [51]. Regarding the serovars associated with seafood, Hadar is frequently reported in Latin America and Africa. Typhimurium and Senftenberg are frequently reported in Europe. Weltevreden and Newport are most frequently reported in Asia and North America, respectively [51].

#### ***1.2.2.4 Transmission modes of Salmonella***

The main reservoirs of *Salmonella* are the intestinal tract of humans and animals, particularly wild birds and reptiles [4]. *Salmonella* is primarily transmitted through consumption of contaminated food including animal-based foods, vegetables or fruits [2]. An estimated 55% (range 32-88%) of human NTS cases are due to consumption of contaminated food worldwide [47,53]. Approximately 9% (range 0-19%) of human NTS infections are acquired through direct animal contact [53] while 13% (range 0-29%) of human NTS infections are attributable to the environmental sources including soil, water and NTS-contaminated animal faeces [45,54,55]. Direct human-to-human transmission accounts for 9% (range 0-19%) of human NTS infections, while 14% (range 3-26%) of human NTS infections are travel-related [45,54,55]. Human-restricted serovars such as Typhi is primarily transmitted human-to-human through faecal contamination [56].

#### **1.2.3 Virulence factors**

Virulence factors involved in the pathogenic process of *Salmonella* include protein secretion systems, toxins, fimbriae adhesins, flagella and others [57]. These virulence factors are not equally present in all NTS serovars and contribute to differences between serovars in pathogenicity, virulence and host range [45].

##### ***1.2.3.1 Protein secretion systems***

Protein secretion systems include the Type III secretion system (T3SS), outer membrane vesicles (OMV), and the Type VI secretion system (T6SS) (128-130).

The T3SS is one of the major protein secretion systems and is comprised of a secretion apparatus (20 to 25 structural proteins), regulatory proteins and translocated effector proteins [58-60]. *Salmonella* harbors two distinct T3SSs located on *Salmonella* pathogenicity islands (SPIs) (SPI-1 and SPI-2). SPI-1 T3SS mediates early stages of the

infection, while SPI-2 T3SS acts in systemic infection and modulates host cell signalling cascades to allow *Salmonella* proliferation [61-64].

OMVs are released from the cell surface of Gram-negative bacteria and consists of outer-membrane proteins, LPS, and phospholipids [65,66]. OMV play a role in translation of a subset of T3SSs-independent secreted proteins PagJ, PagK1, and PagK2 into the host cytosol [67]. The pore-forming cytotoxin factor ClyA in Typhi is released via OMV [68,69].

The T6SS represents a recent identified protein secretion in many Gram-negative bacteria and participates in inter-bacterial killing and pathogenesis but not essential for virulence [70-72]. There are five phylogenetically distinct T6SSs encoded by 5 SPIs (SPI-6, SPI-19, SPI-20, SPI-21 and SPI-22) [72,73].

#### **1.2.3.2 Toxins**

*Salmonella* secretes a few exotoxins, including the cytotoxins and the enterotoxins [57]. *Salmonella* cytotoxins are an outer membrane component and may be involved in cell damage and/or invasion [74]. *Salmonella* typhoid toxin, also called cytolethal distending toxin (CDT), was originally found in serovar Typhi and is present in over 41 NTS serovars but absent in worldwide serovars Typhimurium, Enteritidis, and Newport [45,75]. CDT is encoded by *cdtA*, *cdtB* and *cdtC* genes located in cytolethal distending toxin islet. CDT can cause limited DNA damage, thereby controlling the host cell cycle [76]. In addition, an ArtAB toxin is present in Typhimurium DT-104 strains and other NTS serovars [77,78]. ArtAB toxin is encoded within a prophage and is associated with prophage excision [79,80].

*Salmonella* enterotoxin (Stn) is secreted by all *S. enterica* serovars and encoded by *stn* gene located on chromosome [81,82]. Stn may be a virulence factor in the pathogenesis of *Salmonella* and be responsible for the enterotoxicity of *Salmonella*. However, its importance is conflicting [83-85].

### **1.2.3.3 Fimbriae**

Fimbriae (Pili) are proteinaceous surface appendages responsible for attachment and adhesion to the host cell [86]. There are 39 putative fimbrial operons identified in *Salmonella* according to phenotypic and genomic analyses [45].

Three pathways which are the nucleator dependent pathway, the type IV fimbriae and the chaperone–usher-dependent pathway have been described for the assembly of fimbriae [57,87]. The nucleator dependent pathway is encoded by the *agf* (aggregative fimbriae) operon and may be useful for bacterial adhesion and invasion [45,88]. The type IV fimbriae are encoded by the *pil* operon located on SPI-7 [57]. The chaperone–usher-dependent pathway is encoded by the remaining 36 fimbrial operons [45,57].

### **1.2.3.4 Flagella**

Flagella are long helical filaments attached to rotary motors and confer motility of *Salmonella* species [89]. Flagella also induce the host innate immune response [45,90]. The regulation of flagella upon infection may reduce or prevent activation of a host immune response [45,91]. Flagella is observed in most NTS serovars [45,57]. The major subunit (antigenic) of flagella is encoded by *fliC* (H1), *fljB* (H2) [92]. The secretion of structural subunit proteins of flagella is determined by a flagellum-specific T3SS [93].

## **1.2.4 Mobile genomic elements**

### **1.2.4.1 *Salmonella* Pathogenicity Islands (SPIs)**

SPIs are genetic elements located in the chromosome and have been acquired by horizontal gene transfer. SPIs harbor most important virulence genes and other proteins essential for host cell invasion and intracellular pathogenesis [94]. There are 24 SPIs been identified so far, and the majority of SPIs are associated with tRNA genes [45,95,96].

The distribution of 24 SPIs in *Salmonella* differs. SPI-1 is conserved throughout the genus *Salmonella*. While SPI-22 is found in *S. bongori* only, SPI-20 and SPI-21 are found in *S. enterica* subsp. *arizonae* only [45,97,98]. SPI-15 is specific to Typhi and SPI-7 is associated with human-restricted serovars (Typhi, Paratyphi A, Paratyphi B and Paratyphi C). The remaining SPIs are differentially distributed among *S. enterica* [45,99]. The SPIs are summarised in Table 1.2-1.

#### **1.2.4.2 Virulence Plasmids**

Nine *S. enterica* serovars (Typhimurium, Enteritidis, Dublin, Choleraesuis, Gallinarum/Pullorum, Abortusovis, Paratyphi C, Abortusequi and Sendai) are known to carry a low copy-number *Salmonella* Virulence Plasmid (pSLT) containing virulence genes [100-104].

pSLTs are serovar specific which vary in size and genetic content but all harbor the *spv* (*Salmonella* plasmid virulence) operon which consists of *spvRABCD* genes [45,105-107]. The *spv* operon encodes a toxin that alters the host cell cytoskeleton to enhance bacterial survival, thus increasing virulence [108]. Some pSLTs carry plasmid-encoded fimbriae (*pef*) fimbrial operon encoding an adhesive type of fimbria or the conjugal transfer gene *traT* and the uncharacterised *rck* and *rsk*. These additional virulence genes may contribute to other stages of the infection process [107,109].

Typhi carries a plasmid HCM1 belonging to antimicrobial resistance plasmid family, *incHI1* which confers multiple-drug resistance to antimicrobial agents and heavy metals [97,110,111].



**Table 1.2-1: Summary of *Salmonella* pathogenicity islands (SPIs)**

<b>SPI</b>	<b>Location</b>	<b>Main encoding genes</b>	<b>Roles in pathogenesis</b>	<b>References</b>
SPI-1		<i>invA</i> , T3SS and effector proteins	Invasion of epithelial cells and macrophage apoptosis	[108,112]
SPI-2	tRNA <i>valV</i>	T3SS and effector proteins	Intracellular survival, replication in both epithelial cells and macrophages	[113,114]
SPI-3	tRNA <i>selC</i>	Magnesium transport system ( <i>mgtCB</i> )	Intramacrophage survival	[94,115]
SPI-4	tRNA <i>ssb</i>	T1SS ( <i>siiABCDF</i> ), non-fimbrial adhesin ( <i>siiE</i> )	Intramacrophage survival, toxin secretion	[94,116]
SPI-5	tRNA <i>serT</i>	<i>sopB</i> (effectors of SPI-1), <i>pipB</i> (effectors of SPI-2) and <i>pipACD</i>	Epithelial invasion, enteric salmonellosis, and chicken colonization	[94,117]
SPI-6	tRNA <i>saf</i>	T6SS, fimbriae genes ( <i>safABCD</i> ) and invasion <i>pagN</i> gene	Invasion, intramacrophage survival, chicken colonization	[72,94,118]
SPI-7 (MPI)	tRNA <i>pheU</i>	Vi capsule biosynthesis genes, <i>sopE</i> phage	Vi exopolysaccharide and intramacrophage survival	[94,119-122]
SPI-8	tRNA <i>pheV</i>	Bacteriocin fragment	Unknown	[94,97,123]
SPI-9	Lysogenic bacteriophage	T1SS, adhesin and gene STY2875 similar to a large RTX-like protein	Epithelial adherence	[97,124]
SPI-10	tRNA <sup><i>leuX</i></sup>	<i>sef</i> operon <i>sefD</i> encoding P4-like prophage, <i>sef/pef</i> fimbrial	Intramacrophage uptake or survival and virulence in mice and chickens	[125-127]
SPI-11	CdtB-islet	Typhoid toxin gene islet ( <i>cdtB</i> , <i>pltA</i> , <i>pltB</i> )	Typhoid fever pathology	[75,128]
SPI-12	tRNA <i>proL</i>	<i>sspH2</i>	Improvement fitness in the host	[119,129]
SPI-13	tRNA <i>pheV</i>	Genes for enzyme regulators and LysR family transcriptional regulators	Macrophage survival in chickens	[97,99]

SPI-14		6 genes encoding a putative acyl-coenzyme A (Co-A) dehydrogenase	Chicken pathogenicity	[45,117,126]
SPI-15	tRNA <i>glyU</i>	A phage integrase gene and 4 hypothetical protein-coding genes	Unknown	[130]
SPI-16	tRNA <i>argU</i>	Bactoprenol glucose translocases ( <i>gtrAB</i> ) and a phage integrase	LPS modification, seroconversion	[45,130]
SPI-17	tRNA <i>argW</i>	Six ORFs (high homology to genes of SPI-16)	LPS modification, seroconversion	[45,130]
SPI-18		Hemolysin <i>hlyE</i> (known as <i>clyA</i> or <i>sheA</i> ), Typhi-associated invasin A protein ( <i>taiA</i> )	Epithelial invasion, phagocytosis	[75,131,132]
SPI-19		T6SS	Intramacrophage survival, chicken colonization	[73,99]
SPI-20		T6SS	Unknown	[98,99]
SPI-21		T6SS	Unknown	[98,99]
SPI-22		T6SS	Unknown	[99]
SPI-23		T3SS effectors, T4SS pilin protein ( <i>potR</i> and <i>talN</i> )	Host cell adherence and invasion, invasion of pig epithelial cells and tissue tropism	[95,96]
SPI24	CS54	Outer membrane protein ( <i>shdA</i> , <i>sivH</i> , <i>ratAB</i> , <i>sinI</i> )	Fibronectin binding, adherence/invasion of fibronectin-producing cells	[97,133,134]

T3SS: Type III secretion system; T1SS: Type I secretion system; T6SS: Type VI secretion system; MPI : major pathogenicity island; ORF: open reading frame; CdtB: cytolethal distending toxin subunit B; CS54: genetic island located at centisome 54

#### 1.2.4.3 Prophages

*Salmonella* genomes contain several prophages and prophage remnants [97,135]. The prophages associated with T3SS translocated effector proteins include P2-like prophage SopEΦ, 3 lambda-like Gifsy phages and phage remnants [135,136].

The P2-like prophage SopEΦ is located outside of SPI-1 and contains the gene *sopE* encoding the SPI-1 T3SS translocated effector protein SopE, which is involved in the invasion and inflammation of host cell through activating the host cell RhoGTPases Cdc42 and Rac1 [61,135,137,138]. SopEΦ is present in several isolates belonging to *S. enterica* subspecies I as well as *S. enterica* subspecies IV and VII [139]. However, not all *sopE* positive isolates harbor a P2-like phage SopEΦ. Some *sopE* gene positive isolates harbor a cryptic lambda-like phage similar to Gifsy-phages encoding the conserved *sopE* gene cassette (*sopE*-moron) [61,140]. This demonstrated that lysogenic conversion with SopEΦ or Gifsy phages can cause the transmission of additional genetic material encoding effector protein between *Salmonella* strains [61,140-142].

Of 3 lambda-like Gifsy phages, Gifsy-1, Gifsy-2 and Gifsy-3, phage Gifsy-1 is integrated into the 5' end of the host *lepA* gene and encodes the effector protein GogB of YopM family (leusine-rich repeat protein) [135,141,143]. Gifsy-1 carries a potential virulence modulating gene *gipA* which is specifically involved in the bacterial colonization of the small intestine [135,144].

Phage Gifsy-2 is integrated between *pncB* and *pepN* and encodes the SPI-2 T3SS effector protein SseI (also termed GtgB or SfrH) [135,143,145,146]. Gifsy-2 also carries the gene *sodCI* (periplasmic superoxide dismutase) encoding a periplasmic Cu/Zn superoxide dismutase and the gene *gtgE* [135,147-149]. Together, *gtgE* and *sodCI* are responsible for the potential virulence of Gifsy-2 [147].

Phage Gifsy-3 carries the phoP/phoQ-activated *pagJ* gene and *sspH1* gene encoding T3SS effector protein SspH1 (leucine-rich repeat protein), which modulates the production of intestinal epithelial cell invasion [141,146,150-152].

Phage remnants (bacteriophage-like sequences) encode two further T3SS effector proteins SopE2 and SspH2 [58,136]. The effector protein SopE2 is an activator for Cdc42 and shows 69% sequence similarity to SopE [153]. While effector proteins SopE2 co-localizes with vacuole-associated actin polymerizations (VAP) to reduce or remodel VAP [154].

#### **1.2.4.4 The tRNA<sup>leuX</sup> island**

The chromosomal locus tRNA<sup>leuX</sup> island of *Salmonella* encodes SPI-10 and contains many P4 phages, plasmid and transposable element-related genes or gene fragments, suggesting that the tRNA<sup>leuX</sup> island is a hypervariable region associated with horizontal gene transfer across the *Salmonella* genus [125]. The genes in the tRNA<sup>leuX</sup> island are different between serovars and within serovars [125,155].

The tRNA<sup>leuX</sup> island of Typhimurium carries genes designated STM4488 to STM4498 as locus tags [125,156]. These genes are absent in the majority of serovars except a small number of SARB strains. Derby SARB9 and Stanleyville SARB61 were positive to both STM4493 and STM4496-STM4498. Saintpaul SARB56 contains STM4496-STM4498, STM4492 and STM4495 [125].

#### **1.2.4.5 Enteritidis-specific genomic island**

The Sdf I (*Salmonella* difference fragment) specific to Enteritidis has been identified using suppression subtractive hybridization and is a genomic region of ~4,060 bp located on the chromosome adjacent to the *ydaO* gene [157]. This region contains 6 genes designated *lygA* to *lygF* which have been used as genetic markers for specific detection of Enteritidis [157].

The Enteritidis-specific genome island (GEI) has been identified through *in silico* comparison genomic sequences analysis by Santiago *et al.* [158]. The GEI is an ~12.5-kb segment located at 5' end of gene SEN1377 (*ydaO* or *ttcA*) and harbors annotated 21 genes (SEN1378 to SEN1398) which encode phage-related proteins [158]. The GEI has been designated as the defective prophage  $\phi$ SE14 previously [159]. Sdf I is an internal genomic region of  $\phi$ SE14 and the genes *lygA* to *lygF* in Sdf I correspond to SEN1379, SEN1380, SEN1382, SEN1383 and SEN1384 in  $\phi$ SE14, respectively [158].

### **1.2.5 Detection, identification and serotyping of *Salmonella***

#### ***1.2.5.1 Culture-based methods for detection and phenotypic serotyping***

Conventional culture-based methods for detection and isolation of *Salmonella* include cultures on selective media and characterization of candidate colonies by biochemical tests followed by serotyping. Firstly, pre-enrichment uses a nutritious nonselective medium (BPW and lactose broth) to enhance *Salmonella* growth [160]. Secondly, selective media (BGA, brilliant green agar; BSA, bismuth-sulfite agar; XLD, xylose-lysine-deoxycholate agar) is used for growth of presumptive positive *Salmonella* colonies [161-163]. Then isolated presumptive *Salmonella* colonies from plating media are incubated in triple sugar iron agar (TSI) for isolation of pure cultures. Finally, pure cultures are examined by morphological and biochemical tests for identification and confirmation of *Salmonella* and agglutination reactions for serotype identification [160].

Traditional phenotypic serotyping by slide agglutination of the *Salmonella* isolate with specific polyvalent antisera is performed to identify variants of O and H antigens. Serotype is then assigned based on the White-Kauffmann-Le Minor scheme [12]. Conventional culture-based methods are considered as useful for food safety and public health surveillance [160] and traditional phenotypic serotyping is the golden standard for *Salmonella* identification and characterization [20,164]. However, isolation and confirmation takes more than 5 days to complete. Cross-reaction can occur in highly similar serovars and strains with partially formed O antigens (mucoïd and rough strains). Furthermore, nonspecific agglutination may cause false positive results and autoagglutination or loss of antigen expression may result in unidentified serovars [20,165,166].

#### ***1.2.5.2 Laboratory methods for molecular serotyping targeting serotype specific O and H antigen genes***

Molecular serotyping including polymerase chain reaction (PCR) based methods and probe-based methods have been developed for *Salmonella* serotyping by directly targeting serotype specific O antigen genes (*wzx* and *wzy*) and H antigen genes (*fliC* and *fliB*) [20,167-172]. In 2007, the combination of 3 multiplex PCRs targeting 5 major O antigens (*wzx* and *wzy*), 8 H1 antigen (*fliC*) and 7 H2 antigens (*fliB*) were developed to

obtain complete serotypes of 423/500 (84.6%) routine isolates [167]. Recently, PCR-based detection of O antigen genes along with amplification the internal variable region of *fliC* and *fljB* provided the best serovar determination [173].

The multiplex bead-based suspension array (Bio-Plex array) developed by Fitzgerald *et al.* could identify 95% of O antigens among 200 isolates based on the O-antigen-encoding biosynthetic *rfb* genes [17]. A probe-based approach was able to correctly differentiate 80% of 36 different H antigens genes among 500 isolates [168]. Several probe-based assays targeting O and H antigens genes have been developed for serotyping [171,172,174].

Molecular serotyping targeting serotype specific O and H antigen genes have provided alternatives for rapid identification of *Salmonella* due to the concordance with traditional serotyping. However, the existing primers or probes in these methods do not cover uncommon serovars or new serovars.

#### **1.2.5.3 Laboratory methods for molecular serotyping targeting genomic markers**

Molecular serotyping methods have been developed for serovar prediction based on proxy or surrogate markers unrelated to the O and H antigens genes, such as virulence genes and serovar-specific genes, DNA fragments or genomic regions and serotype-specific CRSIPR loci that are correlated with serovars [164,175-180].

A multiplex PCR method has been developed by Kim *et al.* targeting 12 genes which were identified in Typhimurium LT2 and Typhi CT18 from whole-genome sequence comparisons [123,175,181-183]. This method can serotype 30 of the most common clinically relevant serovars based on the presence or absence of 12 genes and had 97% accuracy of serovar prediction [175]. In 2010, Peterson *et al.* developed a multiplex PCR assay called *Salmonella* Typing Virulence (STV) by adding 3 additional virulence genes *spvC*, *invA* and *sseL* and 2 genes targeting Typhimurium and Enteritidis into Kim's multiplex assays [176]. All *Salmonella* serovars carry the *invA* gene and 4 serovars including Typhimurium, Choleraesuis, Dublin and Enteritidis harbor the *spvC* gene [184,185]. This STV multiplex PCR was able to predict 42 serovars and the accuracy was determined to be 95.3% (135 of 142 isolates) [176].

Several real-time PCR assays targeting genes *sefA*, *sdf*, *aceK*, *sdr*, *spv* and *floR* have been developed for prediction of common serovars [186,187]. Arrach *et al.* developed a real-time PCR which was able to correctly identified 12 serovars on the basis of the presence and absence of 146 genes obtained from comparative genome hybridization on 291 *Salmonella* isolates representing 32 serovars [164].

#### **1.2.5.4 Sequence-based molecular subtyping method**

MLST is a sequence-based molecular subtyping method for population genetic analyses of pathogenic microorganisms [188]. The *S. enterica* MLST schemes include classical seven-gene MLST, ribosomal MLST (rMLST), core genome MLST (cgMLST) and whole genome MLST (wgMLST) [22,188-192]. The classical seven-gene MLST assigns *Salmonella* strains into sequence type (ST) based on sequence comparisons of 7 housekeeping genes [22]. STs often correlate with serovars and thus serovar is predicted using the ST matches in the MLST database for a query strain [22].

#### **1.2.5.5 WGS based *in silico* serotyping**

WGS facilitates accurate *in silico* *Salmonella* serotyping [193-195]. Three novel Web-based platforms, *Salmonella in silico* Typing Resource (SISTR), SeqSero 1 and SeqSero 2, are available for rapid prediction of *Salmonella* serovars using WGS data [193-195]. The SISTR is a platform for rapidly serotyping *Salmonella* genome assemblies [193]. The SISTR platform utilizes the sequences of O and H antigen genes and/or genoserotyping serogroup-specific probes to predict the serovar according to the *Salmonella* antigenic formula. To refine genoserotyping prediction, MLST [22], rMLST [190] and cgMLST [191,192] are incorporated into the SISTR platform. In addition to serotype determinants, markers derived from cgMLST based phylogenies enhance the overall serovar prediction accuracy to over 94.6% on validation of 4,291 *Salmonella* genomes [193].

SeqSero is another platform for *Salmonella* serovar prediction using raw sequencing reads or genome assemblies [194,195]. SeqSero has two versions which are original SeqSero (SeqSero 1) [195] and SeqSero2 [194]. SeqSero can predict more than 2,200 serovars based on O antigen determinants of the *wzx* or *wzy* genes and H antigens determinants of the *fliC* and *fliB* genes [194,195]. In addition to SeqSero 1, additional

markers for minor O-antigen epitopes and for subspecies identification are used in SeqSero2. Furthermore, using k-mer-based algorithm, SeqSero2 increases the accuracy of prediction for genome assemblies from 86.5% (SeqSero1) to 94.1% [194]. Compared to SeqSero 1, SeqSero2 improved the accuracy for overall serovar prediction from 95% to 98% and reduced the multiple serovar prediction rate from 33% to 13% for both raw sequencing reads and genome assemblies [194,196].

rMLST is a curated MLST scheme by indexing variations of 51 ribosomal protein-encoding genes (*rps*) [190]. The ribosomal sequence type (rST) can provide rapid taxonomy and typing, enabling the interpretation of the extensive diversity within *Salmonella* [23,190]. These are high resolution methods that will cluster isolates with other isolates of the same serovar therefore providing serotyping in an indirect way. cgMLST is a gene-by-gene approach focusing on core genes for genome-based phylogenetic analysis [191,192]. *Salmonella* cgMLST and wgMLST are the schemes for *Salmonella* characterization and epidemiological tracing [23]. *Salmonella* rMLST, cgMLST and wgMLST have been implemented in a web-based platform EnteroBase [23].

#### ***1.2.5.6 Rapid, accurate and sensitive detection of Salmonella using laboratory diagnostic methods***

Nucleic acid amplification assays such as PCR based methods (multiplex PCR, real-time PCR) and isothermal methods [loop-mediated isothermal amplification (LAMP) and multiple cross displacement amplification (MCDA)] are the most common methods for rapid detection of *Salmonella* without requiring culture and isolation [197-199]. However, only a few assays are available for serovar detection and identification [200-208].

Among these assays, the MCDA assay is a novel isothermal strand-displacement polymerization reaction technique devised by Wang *et al.* in 2015 [198]. The MCDA assay employs 10 sequence-specific primers (6 primers in LAMP and 2 primers in PCR) to amplify the target. Ten primers consist of cross primers (CP1 and CP2), 2 displacement primers (F1 and F2) and 6 amplification primers (D1, C1, R1, D2, C2 and R2). Ten primers binding ten distinct sites in the MCDA assay facilitate its specificity and sensitivity. Moreover, the MCDA assay is easy to perform and can obtain quick results in a short time (about 40-min) [198]. Compared with LAMP, the sensitivity of MCDA



increased by 16-fold and the positive reactions of MCDA was 10 minutes faster than the LAMP [198].

The MCDA assay has been used to detect *Salmonella* by targeting *invA* gene at species level [197]. The *Salmonella* MCDA detected 6.25 fg pure DNA templates per reaction and observed the positive reactions in as little as 12 min [197]. Compared with *Salmonella*-qPCR, *Salmonella* MCDA experienced at least 400-fold increase in sensitivity and approximately 20 minutes faster for observation of positive results for pure culture [197].

### **1.3 *Shigella***

*Shigella* is gram negative, facultative anaerobic, rod-shaped intracellular bacterial pathogen. *Shigella* is responsible for the third most common foodborne diarrhoea disease caused by bacterial pathogens and the second-leading cause of diarrhoeal deaths worldwide [209,210]. *Shigella* can cause human shigellosis varying from mild diarrhea to bacillary dysentery (severe mucoid/bloody diarrhea ) via an exceptionally low infectious dose (<10 cells) [211]. The estimated *Shigella* infections is 188 million cases annually, resulting in 160,000 deaths predominantly in young children [210].

*Shigella* is closely related to *E. coli*, a commensal microflora in humans and warm-blooded animals with different pathotypes that causes a range of diseases [212,213]. Pathogenic *E. coli* is classified into non-enteroinvasive *E. coli* and enteroinvasive *E. coli* (EIEC) [213]. EIEC can cause bacillary dysentery in humans using the same invasive mechanisms as *Shigella* [214]. Due to their similarity, the differentiation of *Shigella* from EIEC is important for clinical diagnosis and public health epidemiologic investigations.

#### **1.3.1 Taxonomy and classification**

##### **1.3.1.1 *Shigella***

*Shigella* is named after the Japanese bacteriologist Kiyoshi Shiga, who first discovered it in 1897 [215]. The genus *Shigella* belongs to the family Enterobacteriaceae and is divided into four species including *Shigella flexneri*, *Shigella sonnei*, *Shigella boydii* and *Shigella*

*dysenteriae* based on biochemical and serological typing. *Shigella* species are further divided into serotypes according to O-specific polysaccharide (O antigen) of the LPS [216]. *S. dysenteriae*, *S. flexneri* and *S. boydii* consist of 15 serotypes, 19 serotypes, 20 serotypes respectively. Whereas *S. sonnei* possesses a single serotype. In this thesis, SF, SS, SB and SD are the abbreviation of *S. flexneri*, *S. sonnei*, *S. boydii* and *S. dysenteriae* respectively.

*Shigella* has O antigens and lacks H antigens [217]. *Shigella* serotypes are determined by variation in the genes *wzx* and *wzy* that encode the O antigen [218]. There are 35 *Shigella* distinct O antigens for *Shigella* serotyping [218]. The SD and SB serotypes are designated 1 to 15 and 1 to 20, respectively. The abbreviation of “species” name plus the serotype number will be designated to a serotype throughout this thesis (e.g. *S. dysenteriae* serotype 1 it will be referred to SD1).

SF serotyping is complicated since all SF serotypes except SF6 share the same O antigen polysaccharide backbone [219]. To identify all non-SF6 SF serotypes, a series antigenic determinants (O-factors) classified as either type or group are involved in SF serotyping [219,220]. O-factors are encoded by O antigen modification genes including a glycosylation (*gtr*) operon, O-acetylation (*oac*) genes located on bacteriophages as well as O antigen phosphoethanolamine transferase (*opt*) genes on plasmid [222, 225, 226]. Roman numerals I, II, III, IV, V, VI, and VII are used to define the type of O-factors and Arabic numerals 3,4; 6; 7,8; 9; and 10 are used to define the group of O-factors [219,221]. The SF serotypes are assigned by a combination of O antigen and O-factors [220].

Molecular evidence indicates that the genus *Shigella* and *E. coli* belong to the same species [222,223]. In the 1940s, *Shigella* was formally recognised as a genus separated from *E. coli* [13]. Additional genetic analyses indicated that *Shigella* is a metabolically inactive biotype of *E. coli* [224]. However, the genus *Shigella* consists of four species SD, SF, SB and SS in the current classification scheme, corresponding to subgroup A, subgroup B, subgroup C, and subgroup D, respectively, based on the Congress of the International Association of Microbiologists in 1950 recommendation [225].

### **1.3.1.2 Enteroinvasive *E. coli* (EIEC)**

EIEC has been originally reported as ‘paracolon bacillus’ in 1944 and shigellosis-like symptoms caused by EIEC was first shown in 1971 [226]. EIEC is classified into serotypes according to *E. coli* Kauffmann serotyping scheme [227].

*E. coli* has three antigens: the highly polymorphic somatic O antigen, flagellar H antigen and capsular K antigen. The O antigen is composed of two to seven oligosaccharide repeating O units [228,229] and are encoded by O-antigen biosynthesis genes located on the O antigen gene clusters [230,231]. H antigen is determined by the flagellin which is encoded by *fliC* and some additional flagellin gene (*fliA*, *fliB*, *fliC* or *fliD*) [232-238]. The current *E. coli* serotyping scheme has 188 O antigens designated O1 to O188 and 54 H antigens designated H1 to H56 [230,238]. Notably, 6 of 188 O antigens and 3 of 56 H antigens have been withdrawn [239-241]. The K antigen is a capsular polysaccharide antigen present in a proportion of *E. coli* strains with over 80 types [242]. K antigen are co-expressed with one of O8, O9, O20, or O101 groups [243]. The variation in O units provide the major basis for the serotyping schemes [229]. The *E. coli* serotype is designated according to the specific combination of O and H antigens.

EIEC are assigned to 24 *E. coli* serotypes (O28ac:H-, O29:H-, O112ac:H-, O115:H-, O121:H-, O124:H-, O124:H7, O124:H30, O124:H32, O135:H-, O136:H-, O143:H-, O144:H-, O144:H25, O152:H-, O159:H-, O159:H2, O164:H-, O167:H-, O167:H4, O167:H5, O173:H-, and recently O96:H19 and O8:H19) [244,245]. Only a few EIEC have the H antigen [246] and some EIEC O antigens are similar or identical to the typical *Shigella* O antigens, such as O112ac (SD2), O121 (SD7), O124 (SD3), O143 (SB8), O152 (SD12), and O167 (SB3) [218].

## **1.3.2 Epidemiology**

### **1.3.2.1 *Shigella***

#### **1.3.2.1.1 Global prevalence**

The epidemiology and geographical distribution of *Shigella* varies between the four *Shigella* species and their various serotypes. *Shigella* is the third leading cause of bacterial foodborne diseases globally and occurs predominantly in sub-Saharan Africa and South Asia [247,248]. SF is prevalent in developing countries in sub-Saharan Africa as well as

parts of Asia and is responsible for up to 62% of all cases of *Shigella* infections. In contrast, SS is prevalent in economically transitional states or developed countries and accounts for up to 80% of all cases of *Shigella* infections in these regions [249]. SB infections are most common in Bangladesh and South-East Asia and uncommon outside of these regions [250]. SD is rarely isolated in current surveillance [209]. SB and SD cause less than 5% each of all cases of *Shigella* infections globally [210].

Among *Shigella* serotypes, SF2a is the most prevalent serotype associated with human bacillary dysentery worldwide [251], SF3a and SF1a are the second and third most prevalent serotypes in Asian countries respectively [252]. SD1, the first identified member of the genus *Shigella*, is responsible for the epidemics and pandemics of severe Shiga dysentery in all age groups and in the developing countries, particularly in Africa [253].

#### ***1.3.2.1.2 High risk population groups***

*Shigella* is the most prevalent pathogen causing moderate to severe diarrheal disease among children 24 to 59 months old and is responsible for endemic diarrhoeal disease among children 1 to 4 years old living in developing countries [210,248]. The incidence of shigellosis in children younger than 5 years old in Asia is 13.2 cases per 1,000 children per year [249,254]. *Shigella* is also a leading cause of death associated with bacterial diarrheal disease among adults aged 15 to 99 years [255]. *Shigella* is frequently detected in travellers returned from endemic areas or men who have sex with men (MSM) in developed countries [256,257].

#### ***1.3.2.1.3 Reservoirs and transmission modes***

*Shigella* is a highly human-adapted bacterial pathogen, although infections in monkeys and gorillas have been reported [258,259]. *Shigella* is transmitted through the faecal-oral route and direct person-to-person contact. The faecal-oral route is caused by six main sources, contaminated food, faeces, fingers, flies, fomites and contaminated water [260]. *Shigella* can also be transmitted through anal sex and oral-anal contact linked with MSM among predominantly human immunodeficiency virus (HIV)-positive men [261,262].

### **1.3.2.2 EIEC**

EIEC is a human pathogen of bacillary dysentery occurring worldwide. It is very common in both adults and children in low-income countries and is travel-related in high-income countries [245,263]. Very little research on the global disease burden or epidemiology of EIEC have been conducted due to frequent misidentification of EIEC as *Shigella* [264,265].

EIEC causes sporadic cases and some outbreaks. Recently outbreaks and sporadic cases caused by the same a rare EIEC serotype O96:H19 have been reported in Europe including one in Italy involving 109 cases of infection in 2012 and two in the United Kingdom involving 157 cases of infection in 2014 [266-268]. More recently, a confirmed outbreak of EIEC caused by EIEC serotype O8:H19 has been reported in the United States involving 52 cases of infection in 2018 [244].

Humans are the major reservoirs and faecal-oral route as well as direct person-to-person contact are potentially transmission route for EIEC [245,263]. The main sources of EIEC infections are from contaminated food or water. EIEC cases are more likely to be returned travellers from high-incidence areas and less likely to be MSM in developed countries [269,270].

## **1.3.3 The close relationships between *Shigella* and EIEC**

### **1.3.3.1 Phenotypic and biochemical characterization**

*Shigella* are generally nonmotile, lysine-decarboxylase (LDC) negative and lactose negative with the exception of some strains belonging to SS which are late-lactose-fermenting (ferment lactose upon extended incubation) [271-273]. EIEC share similar phenotypic and biochemical properties to *Shigella*, however some strains belonging to a few EIEC serotypes are motile and lactose fermenting [272].

### **1.3.3.2 Genotypic characterization**

*Shigella* and EIEC share a specific plasmid with various names (pWR100 in SF5, pMYSH6000 in SF2a, and pSS120 in SS), but generally termed *Shigella* virulence plasmid (pINV) [274]. The pINV plasmid is as large as ~220 kb and has a conserved 30 kb entry region [274,275]. This region encodes the T3SS apparatus and T3SS effectors



#### **1.3.3.3 Virulence of *Shigella* and EIEC**

*Shigella* and EIEC strains invade mucosal epithelium cells of the large intestine. The virulence factors responsible for this invasion are the T3SS apparatus and T3SS effectors encoded by virulence genes on the entry region of the pINV plasmid [281]. These virulence genes form a cluster of 38 genes in the *mxi-spa-ipa* operon (Figure 1.3-1). The virulence effector proteins can be translocated into the host cell cytoplasm through T3SS to destroy colonic tissue and manipulate the immune response of the host [282,283].

Chromosomal virulence genes are also important in *Shigella*. There are several *Shigella*-specific pathogenicity islands (SHI PAI) containing genes encoding additional virulence factors such as factors involved O-antigen conversion and antibiotic resistance [284].

SD1 is well known for producing Shiga toxin which is encoded by an *stx*-encoding phage located on the chromosome [285]. However, recent studies confirmed that some clinical isolates from SS and SF also contain a new *stx*-encoding phage [286-290]

#### **1.3.3.4 Phylogenetic relationships**

The *Shigella* strains fall into three main clusters (C1, C2 and C3) and five outliers (SS, SD1, SD8, SD10 and SB13) based on *Shigella* phylogenies inferred from sequence variation analysis of eight chromosomes housekeeping genes by Pupo *et al.* [291]. The main clusters all contain a mixing of *Shigella* species across phylogenetic clusters. C1 contains the majority of SB and SD serotypes (SB1–4, SB6, SB8, SB10, SB14, and SB18; and SD3–7, SD9, and SD11–13) plus SF6. C2 contains seven SB serotypes (SB5, SB7, SB9, SB11, SB15, SB16, and SB17) and SD2. C3 contains SB12 and all SF serotypes except for SF6. The three clusters and five outliers are nested within commensal *E. coli* lineages except for SB13.

SB13 is distantly related to *E. coli*. A further analysis of 23 housekeeping gene sequences by Yang *et al.* showed a similar phylogenetic conclusion from analysis of 8 housekeeping gene sequences [292]. Additional studies found that SB13 is not invasive and is highly divergent from *E. coli* and other *Shigella* serotypes [293]. A DNA relatedness study conducted by Hyma *et al.* revealed that SB13 is closely related to *Escherichia albertii*

and formed a distinct *E. albertii*-SB13 lineage and were referred to as the typical SB13 [294]. However, a subset of SB13 that were similar to *Shigella* and *E. coli* and express the SB13 antigen were named atypical SB13 [294]. SB13 lacks the pINV plasmid and *ipaH* gene [294].

EIEC strains fall into four clusters (C4, C5, C6 and C7) with one outlier based on the phylogenies of 32 EIEC strains representing 12 EIEC serotypes inferred from sequence variation analysis of four chromosomes housekeeping genes and two plasmid genes by Lan *et al.* [272]. Of 4 EIEC clusters, C4 has O28, O29, O124, O136, and O164 serotypes, C5 contains O124, O135, O152, and O164 serotypes, C6 has O143 and O167 serotypes, while C7 consists of one serotype O144 [272].

### **1.3.3.5 Evolution of *Shigella* and EIEC**

*Shigella* and EIEC have both evolved from commensal *E. coli* and form the distinctive *Shigella*/EIEC pathovar [272,274,291,295]. The evolution of *Shigella* and EIEC involved the acquisition of the pINV plasmid through horizontal gene transfer and the loss of pathways specific to catabolic and motility, and acquisition of new O antigen gene clusters or modification genes [272,274,291,295].

The existence of 3 clusters and 5 outliers of *Shigella* indicated that *Shigella* had emerged at least seven separate times from commensal *E. coli* by acquisition of the pINV plasmid [223,291]. Five outliers excluding the divergent SB13 must be relatively recent lineages after obtaining the pINV plasmids [218]. Comparative genomics on housekeeping genes also indicated that EIEC evolved from multiple lineages of commensal *E. coli* by convergent evolution, and have emerged more recently than *Shigella* [291,292]. These evolutionary findings are supported by recent WGS based phylogenetic studies [216,284,296].

The highly virulent EIEC serotype O96:H19 which arose recently is an example of the emergence of new EIEC from commensal *E. coli* by acquisition of the pINV plasmid [268]. This event demonstrates the possibility of new EIEC serotypes emerging in the future [268].



### **1.3.4 Detection and identification of *Shigella* and EIEC**

#### **1.3.4.1 Differentiation of *Shigella* and EIEC from non-enteroinvasive *E. coli***

*Shigella* and EIEC are closely related to other *E. coli*. Differentiation of *Shigella* and EIEC from non-enteroinvasive *E. coli* is vital for surveillance. Traditional differentiation methods are based on biochemical tests from conventional bacterial culture [297]. *Shigella* and EIEC are unable to ferment lactose, lack motility and are negative to LDC, although a few exceptions exist [213].

PCR-based molecular methods can be used to differentiate *Shigella* and EIEC from non-enteroinvasive *E. coli* by the targeting genetic marker: the *ipaH* gene [216,298-301]. Culture-independent testing (CIT) methods to detect the *ipaH* gene in faecal samples can also be used for identification of *Shigella* and EIEC [302,303].

#### **1.3.4.2 Differentiation of *Shigella* from EIEC**

Differentiation of *Shigella* from EIEC is made difficult by both sharing similar biochemical and genetic properties. There are only a few biochemical properties including mucate fermentation and/or sodium acetate utilization that differentiate of *Shigella* and EIEC [265,272]. While *Shigella* are negative for mucate and acetate, some EIEC are positive for one or both [265,272]. These biochemical properties can only distinguish some of, but not all EIEC from *Shigella*.

Serval studies developed molecular methods including PCR-based assays by targeting genetic markers for identification of *Shigella* and EIEC [216,284,302,304-306]. A duplex real-time PCR targeting *uidA* ( $\beta$ -glucuronidase) and *lacY* (lactose permease) genes was developed by Pavlovic *et al.* for differentiation of *Shigella* and EIEC based on the gene *uidA* present in both and *lacY* only present in EIEC [305]. This *lacY-uidA* assay can detect *Shigella* or EIEC correctly because the *uidA* gene is only present in *E. coli* and *Shigella* while the *lacY* gene is also present in other *Enterobacteriaceae* [305]. However, SS and SD1 contain the *lacY* gene and some EIEC lack the *lacY* gene [304,307].

Recently, a multiplex PCR assay was developed by targeting “clade-specific marker” combined with *ipaH3* gene for differentiation of *Shigella* [216]. However, the accuracy was later questioned [296,302]. The most recent multiplex PCR assay developed by

Dhakal *et al.* was able to separate EIEC from *Shigella* in cultures from CIT *ipaH*-positive samples using a set of genomic markers identified from comparative genomics [302]. This PCR assay can differentiate EIEC from *Shigella* on the basis of the presence of at least two of six genomic markers and provide subtype EIEC isolates [302]. However, these genomic markers were not extensively tested and were identified using a small number of genomes [302].

Separation of *Shigella* and EIEC has been investigated with unique SNPs. Pettengill *et al.* identified single nucleotide polymorphism (SNP) markers for identification of *Shigella* and EIEC [296]. The SNP markers may have the potential for development of a screening assay. However, these markers may detect other *E. coli*.

#### **1.3.4.3 Differentiation between *Shigella* species**

*Shigella* species disproportionately infect young children in low-resource settings [210]. Differentiation between *Shigella* species is important for clinical and epidemiological investigations. Utilisation of mannitol and decarboxylation of ornithine are used for differentiation of *Shigella* species [295]. Ornithine is decarboxylated only by SS and SB13 while mannitol is decarboxylated by SS, SF and SB. Therefore, biochemical properties are only able to separate SD from SS, SF and SB.

A multiplex PCR assay was developed to differentiate between SS and SF by using the markers associated with *she* PAI [308]. This assay may be able to differentiate between SS and majority of SF serotypes which are the most frequent *Shigella* isolates [308]. Recently, a novel *Shigella* multiplex PCR was designed for differentiation of the four *Shigella* species SS, SF, SD and SB by detecting genetic markers identified by comparative genomics [309]. This assay was tested with only one EIEC strain and limited *Shigella* strains.

A *kmer* based WGS identification approach enabled differentiation of *Shigella* to the species level [310]. However, some EIEC stains were misidentified as SF or SD by this method [310]. SNPs that are found in highly conserved core genes were utilised to construct a hierarchical SNP-based genotyping scheme for identification of SS subtypes.

This WGS-based genotyping scheme can facilitate SS surveillance but is not designed for the initial identification of SS [249].

### **1.3.5 Serotyping of *Shigella* and EIEC**

#### ***1.3.5.1 Traditional phenotypic serotyping***

Traditional phenotypic serotyping of *Shigella* depends upon slide agglutination reaction of serotype specific O antigens and O-factors with various monovalent and monoclonal antisera specific to each serotypes [311,312]. Serotyping of EIEC is determined by agglutination reactions with panels of rabbit antisera based on *E. coli* serotyping scheme [227,239].

Traditional phenotypic serotyping is the current gold standard method for *Shigella* and EIEC determination. On the other hand, the phenotypic serotyping is laborious, expensive and time-consuming. Additionally, cross-reactions can occur and lack of available antisera or loss of antigen expression can lead to inaccurate or incomplete serotyping [313].

#### ***1.3.5.2 Molecular serotyping of *Shigella****

Molecular serotyping methods, including DNA microarray, Restriction fragment length polymorphism (*rfb*-RFLP) and multiplex PCR, directly target O antigen specific biosynthetic genes and modification genes for serotyping.

The first DNA microarray was developed by Li *et al.* in 2009 by targeting 34 distinct *Shigella* O antigen specific genes for serotyping of *Shigella* [314]. The *rfb*-RFLP method was developed for serotyping of *Shigella* by amplifying the O-antigen gene cluster combined with restriction enzyme digestion [315-320]. These methods allow serotyping of nonagglutinating, nontypeable or new serotypes of *Shigella* [315,320]. Recently, multiplex PCR assays were developed for molecular serotyping of SF and were able to determine the SF serotypes by targeting the SF O-antigen synthesis genes and modification genes [321,322]. Nevertheless, these methods cannot type all *Shigella* serotypes [315,320-322]. MLST can assign most *Shigella* to a serotype, however some STs contained multiple serotypes [310].

#### 1.3.5.3 Molecular serotyping targeting EIEC serotype specific O and H antigen genes

*E. coli* molecular serotyping methods can be used for EIEC serotyping, such as PCR assays and microarrays for detection of O antigen genes as well as RFLP analysis of O-antigen gene clusters [238,323-325].

An *E. coli* O-genotyping PCR was developed by targeting O-antigen processing genes *wzx*, *wzy*, *wzm*, or *wzt* and can identify almost all known *E. coli* O serogroups [323]. An RFLP assay was designed by Coimbra *et al.* to amplify *E. coli* O-antigen gene clusters of 148 O serogroups and obtained a unique RFLP patterns from MboII digestion of amplified products for each serogroup [325]. The restriction method (*rfb*-RFLP) can be used for typing isolates that are not typeable by conventional serotyping [325]. Microarrays was also used for *E. coli* O group typing by detection of specific *E. coli* O-antigen gene cluster [238,324]. These methods can only predict O serogroups.

#### 1.3.5.4 WGS based *in silico* serotyping

There are two pipelines, ShigaTyper and SerotypeFinder ,developed for serotyping *Shigella* and *E. coli* respectively from WGS data. ShigaTyper utilizes genetic markers *ipaH\_C*, *EclacY*, *cadA*, and *Ss\_methylase* together with serotype-specific *wzx* and *wzy* genes and modification genes for differentiation of *Shigella* from EIEC and for *Shigella* serotype prediction [326]. However, the *cadA* gene is present in SS and some SD as well as 70% of the EIEC genomes [284,326-328]. The *Ss\_methylase* gene was also present in SD and EIEC serotypes and is associated with bacteriophages [326]. SerotypeFinder utilizes *E. coli* O-antigen genes (*wzx*, *wzy*, *wzm*, and *wzt*) and flagellin genes (*fliC*, *flkA*, *fliA*, *flmA*, and *fliN*) for serotyping of *E. coli* [329].

Compared with conventional serotyping, WGS based *in silico* serotyping pipelines are much more rapid and cost effective, therefore providing an alternative to conventional typing strategies [326,329]. However, not all *Shigella* and EIEC can be serotyped based on O or H type genes from genome sequencing data because the antigen genes in some of strains may not be assembled well or may represent novel type [329,330].

## 1.4 STEC

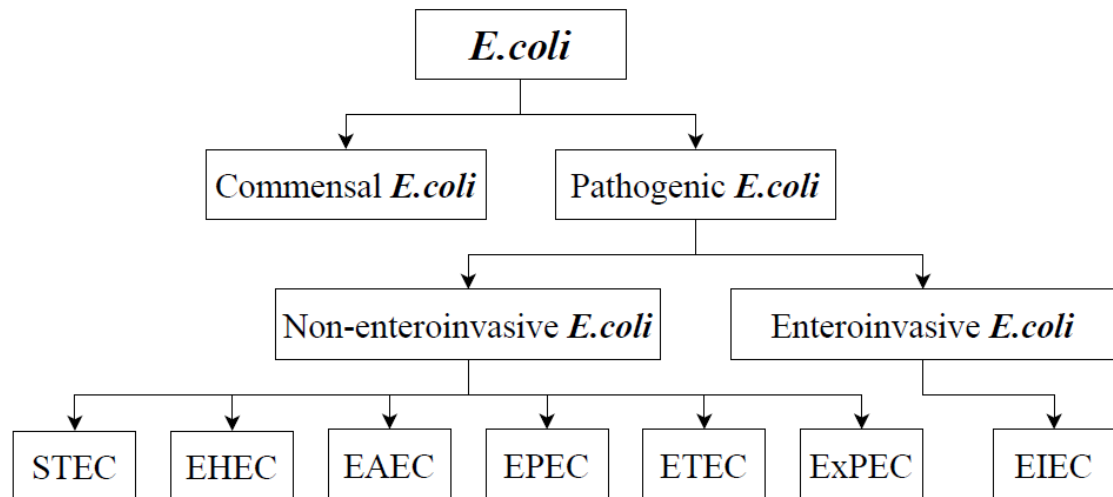
STEC is a pathotype of *E. coli*. As addressed in *Shigella* 1.3 section above, *E. coli* is a commensal microflora as well as a pathogen (non-enteroinvasive *E. coli* and EIEC) with different pathotypes [213]. The pathotypes of non-enteroinvasive *E. coli* are shown in Figure 1.4-1.

STEC is able to produce Shiga toxin and is an important foodborne pathogen in humans worldwide [331]. Highly pathogenic STEC cause diarrheal diseases ranging from mild diarrheal to haemorrhagic colitis (HC) and haemolytic uraemic syndrome (HUS). HUS is the most severe form of STEC infection and can induce acute kidney failure with high proportion of end-stage renal disease (ESRD) and death [331]. The illnesses caused by STEC are estimated to be 2.8 million annually, leading to 3,890 cases of HUS, 270 cases of ESRD and 230 deaths globally [332]. Of these, 1.2 million are considered foodborne, leading to 13,000 cases of DALYs and 128 deaths in 2010 globally [2].

### 1.4.1 Classification, nomenclature and population structure of STEC

The first discovered STEC was in fact EHEC, which is now a subset of STEC and was originally reported from two outbreaks of HC by Riley *et al.* in 1983 [333].

STEC is classified into serotypes according to the traditional *E. coli* Kauffmann serotyping scheme which has been addressed in “1.3.1.2 EIEC” section above [227,334]. There are over 1,100 STEC serotypes listed in a central collection [335]. STEC O157:H7 (EHEC) is the best studied serotype since it has been defined as a human pathogen in 1982 and the remaining STEC serotypes referred as STEC non-O157:H7 serotypes. Additionally, STEC serotypes have been classified into 5 seropathotypes (A through E) by use of the correlation of epidemic and/or serious disease in humans caused by STEC serotypes [336].



**Figure 1.4-1: Pathotypes of *E. coli*.** STEC: Shiga toxin-producing *E. coli*; EHEC: enterohemorrhagic *E. coli*; EAEC: enteroaggregative *E. coli*; EPEC: enteropathogenic *E. coli*; ETEC: enterotoxigenic *E. coli*; ExPEC: extraintestinal pathogenic *E. coli*; EIEC: enteroinvasive *E. coli*.

## 1.4.2 Epidemiology

### 1.4.2.1 Global incidence

STEC is the fourth most common cause of foodborne diseases and foodborne outbreaks worldwide and the estimated global incidence of STEC is 43.1 per 100,000 person-years [332]. Estimated regional incidences ranged from 1.4 per 100,000 person-years in the African sub-regions to 152.6 per 100,000 person-years in the Eastern Mediterranean sub-regions [332]. An estimated 5%–15% of STEC infections develop into HUS and STEC associated HUS is responsible for 2%–5% of mortality from bacterial diarrheal infections [38,332,337]. The highest mortality occurs in American countries (Canada, Cuba, United States) and the South-East Asian subregions. While the lowest mortality occurs in Europe [332].

STEC infection can occur in all age groups but is more frequent and severe in young children [332,338]. Children under 5 years old accounted for 29% of acute STEC infections, 42% of 3,890 cases of HUS, 41% of 270 cases of ESRD, and 29% of 230 cases of deaths annually [332]. The highest incidence of STEC infections in children under 5 years old is 12.2 cases per 100,000 in Argentina, approximately 30-fold higher than in Australia which reports 0.4 cases per 100,000 persons [339,340].

#### ***1.4.2.2 Frequency of diarrhea caused by different serotypes***

STEC has over 1,100 serotypes recognized so far, only 28% of the serotypes are associated with human infections, although many STEC serotypes have only been isolated and reported once in humans [335]. STEC O157:H7 is a leading cause of foodborne infections and HUS in humans worldwide predominately in developed regions such as North America and Europe [341-343]. However, the incidence of STEC non-O157:H7 serotypes associated with foodborne outbreaks and human infections has increased in recent years [48,344].

In Australia, the notification rate of STEC infections was 0.4 cases per 100,000 per year, while the notification rate of STEC O157:H7 infections was 0.12 cases per 100,000 per year from 2000 to 2010 [340]. The estimated overall incidence of HUS caused by STEC was 0.07 per 100,000 per year, and the highest rate of 0.49 cases per 100,000 per year was observed in children under 5 years between 2000 and 2010 [340]. Among those with an identified serotype, O157:H7 accounted for 58% of STEC cases and the remaining 42% was non-O157:H7. The common non-O157:H7 serotypes in human infections were O111 (13.7%), O26 (11.1%), O113 (3.6%), O55 (1.3%), and O86 (1.0%) [340]

In the USA, STEC O157:H7 and non-O157:H7 serotypes account for 63,153 (36%) and 112,752 (64%) cases of annual domestically acquired foodborne STEC infections, respectively [47]. STEC O157:H7 is the major cause of HUS and responsible for 85–95% of cases of HUS in North America [345]. The incidence of STEC O157:H7 was decreased with 0.95 per 100,000 in 2010 compared with 2.17 per 100,000 in 2000 [346,347]. While the incidence of STEC non-O157:H7 serotypes increased from 0.12 to 0.95 per 100,000 and surpassed the national incidence of O157:H7 in 2014 [346,347]. The most prevalent STEC non-O157:H7 serotypes in human infections in the USA are O26, O45, O103, O111, O121 and O145, which are the well-known “Big 6” non-O157:H7 serotypes [348,349]. The “Big 6” non-O157:H7 serotypes were responsible for 83% of STEC non-O157:H7 cases between 2000 and 2010. Accordingly, over 90% of STEC infections are caused by O157:H7 and “Big 6” non-O157:H7 serotypes [346]

In the European Union (EU), STEC caused 7,894 cases of infections, 394 HUS cases of which 272 cases (69.4% ) occurred in the children under 4 years old in 2019 [48]. The

STEC infections in humans caused by STEC O157:H7 decreased from 54.9% in 2012 to 26.6% in 2019. Whereas the non-O157:H7 serotypes infections in humans increased by 28.3% compared with 2012. STEC O26:H11/H- was the most common non-O157:H7 serotype in human STEC infections in Europe, followed by O146, O103, O91 and O145 in the period of 2012 to 2019 [48,350]. The most severe HUS associated with STEC infection are caused by STEC O157:H7 and O26:H11/H- [48].

#### ***1.4.2.3 STEC associated outbreaks***

STEC O157:H7 can cause large foodborne outbreaks worldwide since the first recorded outbreaks of HUS associated with *E. coli* serotype O157:H7 were reported in 1983 [333,351]. In the USA, 390 foodborne outbreaks were caused by O157:H7, resulting 4,928 cases with 1,272 hospitalizations, 299 HUS and 33 deaths between 2003 and 2012 [352]. In the EU, 42 foodborne outbreaks were caused by O157:H7, leading to 273 cases, with 50 hospitalised and one death in 2019 [48]. In Australia, there were 4 outbreaks linked to STEC O157, causing 84 cases and 20 hospitalizations from 2001 to 2009 [340]. The largest reported outbreak of STEC O157 in Australia occurred in Queensland in 2013, resulting in 57 cases [353].

STEC non-O157:H7 serotypes mostly cause sporadic cases except a small number of serotypes such as O26:H11, O45:H2, O103:H2, O111:H8, O121:H19, and O145:NM which caused 674 outbreaks from 1995 to 2017 worldwide, resulting in an average of 14% HUS per STEC outbreak [354,355]. In 2011, two outbreaks of a rare STEC serotype O104:H4 caused more than 4,000 cases and 50 deaths in 16 Europe countries [356-358]. Recently, HUS caused by STEC serotype O80:H2 has been reported in certain regions of France [359].

#### ***1.4.2.4 Transmission***

STEC is a zoonotic pathogen and cattle is the main reservoir although it is also found in other animals [360]. The main sources of infection are contaminated beef products as well as raw milk and other foods [361]. STEC infection is mainly by person-to-person transmission via a direct fecal-oral route. STEC infections can also occur through contact with infected animals [362].



### 1.4.3 Virulence Factors

#### 1.4.3.1 Shiga Toxin (Stx)

Stx is classified into two immunologically distinct types named Stx1 and Stx2, where Stx1 is almost identical to the Stx produced by SD1 [213,363]. The Stx is further classified into subtypes according to the standardized Stx nomenclature based on phylogenetic sequence-based relatedness of the proteins by Scheutz *et al.* [364]. Stx1 has 3 subtypes named Stx1a, Stx1c and Stx1d, and STx2 has 10 subtypes named Stx2a to Stx2k [364-367]. STEC has been identified that can produce up to six Stxs including combinations of Stx1 and/or Stx2 with various subtypes [368,369].

Stxs are potent cytotoxins with AB<sub>5</sub> protein structure containing an enzymatic active A subunit and five identical B subunits [370]. The A subunit possess RNA *N*-glycosidase activity and the B pentamer binds the toxin to globotriaosylceramide or Gb3, which is specific to glycolipid receptor of the target microvascular endothelial surface [371]. The AB<sub>5</sub> protein toxin inhibits eukaryotic protein synthesis and lead to host renal endothelial and intestinal epithelial cell death [372].

Stxs are encoded by the *stx* genes located on Stx-phage [373-375]. Stxs are produced and released into the intestinal lumen through prophage induction and cell lysis, although Stx1 is also expressed due to its own promoter under low iron conditions [376]. The pathogenesis of STEC is varied among STEC strains based on the presence of Stxs [344]. The presence of Stx2 induces more severe diseases (HC and HUS) than Stx1 [377,378]. Among Stx2 subtypes, Stx2a is most highly associated with HUS and increased mortality, followed by Stx2c, and Stx2d<sub>activable</sub> [379-381]. In contrast, Stx2e and Stx2f are less likely to cause human infections but can still be associated with HUS [382-386]. Other subtypes of Stx1 and Stx2 can cause human infections with a milder course of disease [364,377].

Besides STEC and SD1, a small number of SF, SD4 and SS strains also produce Stx1 and a few stains of *E. albertii* can produce Stx2f [286,290,382,387].

#### 1.4.3.2 Intimin

Intimin is another key virulence factor for some of STEC and is an outer membrane adhesin. Intimin is encoded by the chromosomal gene *eae* located in the locus of

enterocyte effacement (LEE) [213]. The adhesin intimin has at least 15 different types and subtypes based on sequence and antigenic variation [388,389]. Of these intimin types, the  $\beta$  (present in some non-O157 STEC) and  $\gamma$  (present in STEC O55:H7 and O157:H7) types are mainly found in human isolates [389-391]. The intimin types have a type-specific variable region [389], which allow the types to be utilized as marker for STEC typing in routine diagnostics and epidemiological investigations [389].

The adhesin intimin has attaching and effacing activity which allow the bacteria attached to intestinal epithelium of the host to cause attaching and effacing (A/E) lesions [392]. The production of intimin is closely associated with severe diarrheal, in particular HUS [388,390].

#### **1.4.3.3 Hemolysin**

The pore-forming EHEC hemolysin (EHEC-Hly) is an active repeats-in-toxin (RTX) cytotoxin and is encoded by the *ehxA* gene (syn. EHEC-*hlyCABD*) located on pO157 plasmid [393-395]. The hemolysin is highly conserved among different STEC serotypes and is responsible for the enterohaemolytic phenotype [395,396]. EHEC-Hly may enhance the growth of STEC towards erythrocytes to release haeme and haemoglobin, leading to extraintestinal complications in humans [397]. The EHEC-Hly encoding region has been used as a diagnostic probe for STEC O157:H7 and other STEC isolates [394].

### **1.4.4 Mobile genetic elements**

#### **1.4.4.1 Stx Prophages**

STEC carry lambdoid Stx prophages (Stx-converting bacteriophages or Stx-phages) that harbour *stx* genes and regulate *stx* gene expression [398]. Previous studies demonstrated that Stx-phages vary significantly in their genetic structure, including sizes and gene compositions [399]. This genetic diversity allows Stx-phages acting as mobile genetic elements to be horizontally transferred into other *E. coli*, leading to the emergence of new STEC [369,398,400]. The carriage of more than one Stx-phage may increase STEC virulence [369,401,402].

The lysis region of Stx-phages harbour the *stx* genes and the phage replication cycle enhances the expression of *stx* genes [403]. The Stx-phages or the *stx* genes may be lost

via induction that occurs during infection and isolation as well as routine subculture [404]. The STEC *stx*-negative isolates, containing similar phenotypic characteristics to those of the STEC isolates and belonging to the same ST, have been reported to cause human STEC infections [404-406]. These *stx*-negative isolates may never acquire Stx-phages or may have lost *stx* genes or Stx-phages. Alternatively, these strains may be the progenitor of STEC lineages before they acquired an Stx prophage [404].

#### **1.4.4.2 Locus of enterocyte effacement island**

In the 1990's, the locus of enterocyte (LEE) island was discovered as a 35-kbp pathogenicity island in EPEC. LEE islands have since been located in STEC O157:H7 and some of STEC non-O157:H7 serotypes as well as in *E. albertii* [213,294,407]. The LEE of STEC encodes genes for a T3SS, intimin (*eae*), the intimin translocated receptor Tir and the Esp effector proteins which are secreted via the T3SS [213,396]. The LEE-encoded virulence proteins form the A/E lesions which enable STEC to adhere to host intestinal epithelial cells [407]. The A/E phenotype is acquired with the LEE through horizontal gene transfer [408,409]. Not all STEC are LEE positive while the LEE negative STECs likely do not carry LEE or have lost the LEE [213].

#### **1.4.4.3 Virulence plasmid O157 (pO157)**

STEC O157:H7 carry a large nonconjugative F-like plasmid designated pO157 [396,410,411]. The pO157 is a highly conserved virulence plasmid [396,410,411] and encodes virulence factors such as type II secretion system apparatus (*etp*), adhesin (*toxB*), hemolysin (*ehxA*), periplasmic catalase-peroxidase (*katP*), *eae* conserved fragment (*ecf*), zinc metalloprotease (*stcE*) and serine protease (*espP*) [396,410,412-416]. The pO157 plasmid may be involved in the adherence to epithelial cells although its role in pathogenesis is unclear [396,417,418].

The pO157 plasmid are present in some non-O157 STEC [396]. The plasmids in STEC O26:H11, O113:H21 and many other STECs share several highly conserved regions with pO157 and also contain important virulence genes [213,417,419].

### **1.4.5 Evolution of STEC and emergence of new STEC serotypes**

#### ***1.4.5.1 Overview***

Horizontal transfer and mobile genetic elements (MGE) have played an important role in the evolution of STEC. The acquisition of MGEs such as *stx* genes or Stx-phage have contributed to the emergence of new STEC, and enhanced STEC pathogenicity [420-422].

#### ***1.4.5.2 Evolution of STEC O157:H7***

STEC O157:H7 strains are unable to ferment sorbitol (non-sorbitol-fermenting) [423], while STEC O157:H- (nonmotile) strains first isolated from an outbreak of HUS reported in Germany in 1988 have the ability to ferment sorbitol rapidly [424]. The sorbitol-fermenting STEC O157:H- cause a higher incidence of HUS in Europe, Australia and Asia and were found to cause more infections in children younger than 3 years [425-427].

STEC O157:H7/H- evolved from the non-toxigenic and less virulent, sorbitol-fermenting EPEC serotype O55:H7 by acquisition of virulence-associated MGE including Stx-phages and pO157 via horizontal gene transfers, followed by the acquisition of O157 O antigen cluster and the loss of O55 O antigen cluster [404,428]. Subsequently, O157:H7 lost the ability to utilize sorbitol to form non-sorbitol-fermenting STEC O157:H7, while sorbitol-fermenting STEC O157:H- was formed by the loss of its motility [404,429,430]. STEC O157:H7/H- may lose *stx* genes from Stx-phages, leading to *stx*-negative O157:H7/H-. However, the loss of the *stx* genes more frequently occurs in sorbitol-fermenting STEC O157:H- than in non-sorbitol-fermenting STEC O157:H7 [427].

#### ***1.4.5.3 Hybrid STEC pathotypes***

The potential acquisition of Stx-phages or/and other virulence genes through horizontal transfer to different diarrheagenic *E. coli* provides the opportunities for the emergence of hybrid STEC pathotypes [365]. The genomic features and virulence factors in hybrid STEC pathotypes are the unique combination of characteristics from multiple pathotypes [431]. These hybrid STEC pathotypes with enhanced virulence pose a high risk to public health [365,422].

An example of a STEC hybrid pathotype causing the devastating STEC outbreak with high morbidity and mortality in Central Europe in 2011 was a STEC/EAEC hybrid of

O104:H4 [356,432]. The STEC/EAEC O104:H4 carries *stx2* genes and a set of virulence genes encoding the aggregative adherence fimbriae [433].

A STEC/ExPEC hybrid of O80:H2 has been identified to be associated with HUS in France in 2016 and is also present in other European countries [359,434-436]. STEC/ExPEC O80:H2 harbors *stx* genes and the virulence genes associated with an ExPEC plasmid (pS88).

More recently, hybrid STEC/ETEC representing rare serotypes (O15:H16, O187:H28, O2:H27, O141:H8, O159:H16, O100:H30, O101:H-, O128:H8 and O136:H12) have caused diarrheal diseases and HUS in humans [422,437,438]. These STEC/ETEC serotypes contain *stx* genes including *stx2e* and ETEC virulence marker *sta* genes [422]. In 2020, a STEC/ETEC hybrid O159:H16 isolated from pigs was identified, which carried a novel Stx2 subtype *stx2k* gene and plasmid-encoded ETEC *sta* gene [365].

#### **1.4.6 Detection, identification and serotyping**

##### ***1.4.6.1 Conventional culture-based: isolation for typing***

The conventional culture-based methods rely on selective and differential media for detection and isolation of STEC [439]. The isolation media for STEC O157:H7 and STEC non-O157:H7 have been developed based on the differential biochemical characteristics. The STEC non-O157:H7, like non-pathogenic *E. coli*, can ferment sorbitol in contrast with non-sorbitol-fermenting STEC O157:H7. Thus, non-sorbitol-fermenting STEC O157:H7 can be separated from sorbitol-fermenting STEC O157:H- and STEC non-O157:H7 using Sorbitol-containing MacConkey agar (SMAC) and its modified antimicrobials agar cefixime-tellurite SMAC [423,439,440]. SMAC medium has high sensitivity (100%), specificity (85%) and accuracy (86%) for isolation of non-sorbitol-fermenting STEC O157:H7 [423,439,440]. In addition to SMAC, CHROMagar™ O157 has been used to detect STEC O157:H7 through a chromogenic substrate [441]. However, SMAC and CHROMagar™ O157 mediums are not suitable for the isolation of STEC non-O157:H7.

To isolate sorbitol-fermenting STEC O157:H- and STEC non-O157, several other chromogenic mediums have been designed, such as CHROMagar™ STEC [442,443].

CHROMagar™ STEC can detect common STEC non-O157 serotypes with the presence of the tellurite resistance (*terB*) gene but also produces high false-positive results [442-444]. Given the diversity of STEC, none of the selective agar is in fact specific for non-O157 STEC [442-444].

#### ***1.4.6.2 Conventional culture-based: confirmation and serotyping***

The presence of Stx in samples can confirm that an infection is most likely caused by STEC. The presence of Stx in the suspected colonies from SMAC were tested by using enzyme-linked immunosorbent assay (ELISA) [445,446]. The Stxs have also been identified directly from clinical samples by immunoassay [447]. ELISA allows all STEC to be detected. However, the detection of Stx is more expensive than culture methods and may yield false-positives or cause misdiagnosis due to the potential loss of the Stx-phages which are unstable [213].

Stx positive samples can also be tested for O and H antigens by *E. coli* latex agglutination [448]. Traditional *E. coli* latex agglutination is slide agglutination with *E. coli* O-specific rabbit antisera according to *E. coli* Kauffman agglutination scheme [227]. The latex agglutination reagents are focussed on STEC O157:H7 and “Big 6” non-O157:H7 serotypes. Therefore it is not useful for other non-O157 STEC or O unidentifiable isolates [449].

#### ***1.4.6.3 Molecular methods for detection of stx genes and other virulence genes***

Molecular methods such as DNA hybridization assays and PCR based assays (conventional PCR, multiplex PCR and real-time PCR) have been developed by targeting *stx1* and *stx2* genes for detection of STEC [450-456]. These assays had high sensitivity and specificity and can assign *stx* to a particular subtype by using subtype specific primers [364,457].

#### ***1.4.6.4 Molecular serotyping targeting E. coli O and H antigen genes***

Molecular methods for targeting *E. coli* O and H antigen genes which have been addressed in detail in “1.3.5.3” section above can be used for STEC serotyping. Recently, numerous PCR based assays were developed for serotyping of O157:H7, “Big 6” non-O157:H7 serotypes and a few common non-O157:H7 serotypes [323,458-468]. These

PCR assays were able to simultaneously serotype one or more clinically relevant STEC serotypes by detection of serotype O antigen and H antigen genes. However, these assays focus on O157:H7 and “Big 6” non-O157:H7 and not all common non-O157:H7 serotypes can be serotyped [323,458-468].

#### ***1.4.6.5 WGS based in silico serotyping***

*in silico* pipeline SerotypeFinder can be used for prediction of STEC serotypes [329,469]. As addressed in 1.3.5 section, WGS based *in silico* serotyping may not predict all STEC serotypes from genome sequencing data due to O or H type genes in some of the strains not very well assembled or represented novel types [329,330].

#### ***1.4.6.6 STEC subtyping in surveillance***

The relatedness of the STEC isolate and the sources of outbreaks have been determined by molecular genotyping methods including pulsed-field gel electrophoresis (PFGE), locus variable-number tandem repeat analysis (MLVA) and MLST [470-474]. PFGE combined with MLVA enhance the discriminative power of PFGE because MLVA can type certain non-typeable strains by PFGE [457,472,475]. STs assigned by seven gene MLST can determine the genetic relatedness between strains and trace the isolates of each ST for global epidemiological investigation [192].

WGS based Single nucleotide polymorphisms (SNP) analysis has been applied to STEC outbreak detection and epidemiological surveillance [476-478]. WGS based *k*-mer analysis has also been used for STEC subtyping [457]. Other methods include wgMLST or cgMLST [457]. The utilization of WGS provides superior discriminatory power for outbreak investigations and for the monitoring of hyper-virulent strains relative to PFGE and MLVA [457,479-481]. However, SNP analysis and *k*-mer analysis also require significant computational infrastructure and expertise which limits their adoption [457,482].

## **1.5 Metagenomic approaches for detection of foodborne pathogens**

Metagenomics can be used as a culture-independent diagnostic method with potential for rapid source tracking of foodborne outbreaks and risk assessment of foodborne pathogens [483-486]. Recent applications of shotgun metagenomics approaches can accurately and rapidly detect *Salmonella* and STEC serotypes in a shorter time period and at a strain-level. These applications have demonstrated the applicability of metagenomics approaches as an alternative to culture-dependent methods [483,484,486-489].

## **1.6 Limitations of existing methods for detection and serotyping of *Salmonella*, *Shigella* and STEC**

Existing methods for detection and serotyping *Salmonella*, *Shigella* and STEC as addressed above have drawbacks. Culture based phenotypic detection and serotyping methods are the gold standard but they are laborious, time-consuming and expensive. Cross-reactions, autoagglutination, nonspecific agglutination, loss of antigen expression and lack of available antisera can lead to inaccurate or incomplete serotyping.

Molecular detection and serotyping methods targeting serotype specific O and H antigen genes do not type all serotypes or types with partial results can occur as well. In addition, while different O types can be distinguished by the presence/absence of O antigen specific genes, different H types can only be differentiated by sequence variation of H antigen gene, making it harder to design molecular assays. The *Salmonella* serovar of an isolate may also be inferred by identifying genes only found in the serovar of interest. This has been performed previously but only for a small number of serovars [164,175-179].

Sequence based MLST can assign STs that often correlates with serotypes and thus most serotypes are predicted using the ST matches in the MLST database for a query strain [22]. However this depends on the target species or serotype. Some STs consist of multiple serotypes. With the development of WGS technology, traditional serotyping is being replaced by molecular and *in silico* serotyping based on O or H type genes from genome sequencing data. However, these methods may predict multiple serotypes.



Untypeable and partial types may be assigned due to the antigen genes in some of the strains not being assembled very well or may represent a novel type.

## **1.7 Comparative genomics of accessory genomes**

### **1.7.1 The accessory genomes of *Salmonella*, *Shigella* and STEC**

The genome of *Salmonella*, *Shigella* and STEC has a core genome and an accessory genome. The core genome is defined as the genes (core genes) which are present in all genomes and the accessory genome is defined as the genes (accessory genes) which are present in some but not all genomes [490]. The accessory genome consist of genes specific to a subset of genomes (subspecies, serotypes) and genes specific to single genome [491].

### **1.7.2 Comparative genomic analysis of accessory genomes for identification of specific genomic markers**

Genomic markers that are present or absent from a strain are a good target for detection and identification either using genomic data or by laboratory diagnostic methods. The major challenge for rapid, highly sensitive and specific detection of foodborne pathogens is to identify highly specific and discriminatory genomic targets [302,309]. Comparative genomic analysis of many available genomic sequences of *Salmonella*, *Shigella* and STEC would provide a powerful application for identification of specific genomic targets, which are especially suitable for development of specific gene markers based diagnostic tools for detection and identification of these foodborne pathogens [302,309,492,493].

## **1.8 Aims of the thesis**

*Salmonella*, *Shigella* and STEC are the common causes of bacterial foodborne diseases worldwide. The burden of these pathogens to the economy and human health is best alleviated by prevention. Early detection and identification of contaminating pathogens forms a key part of this prevention strategy, and can be achieved by detection of highly specific and discriminatory genomic targets. Existing methods for detection and serotyping *Salmonella*, *Shigella* and STEC as addressed above have limitations. To overcome the current issues, this study aims to identify pathogen type-specific gene markers for rapid, highly sensitive and specific identification and differentiation of

*Salmonella*, *Shigella* and STEC either from genomic data or using laboratory diagnostic methods.

*S. enterica* is a highly diverse species with more than 2,000 serovars and the ability to distinguish serovars is vital for public health surveillance. Existing *in silico* serovar prediction approaches utilize surface antigen encoding genes, cgMLST and serovar-specific gene markers or DNA fragments for serotyping. However, these serovar-specific gene markers or DNA fragments only distinguished a small number of serovars [177,178]. **The first aim of this thesis** was to identify serovar-specific gene markers for the most frequent *Salmonella* serovars using the extensive publicly available collection of *Salmonella* genomes. These serovar-specific gene markers can be used for molecular serotyping *in silico* typing of genomic data.

In Australia, more than 85% of outbreaks of human *Salmonella* infections were caused by the five most common *Salmonella* serovars: Typhimurium, Enteritidis, Virchow, Saintpaul and Infantis. Rapid, accurate, and sensitive identification of *Salmonella* serovars is vital for diagnosis and public health surveillance. Recently, an isothermal amplification technique MCDA has been employed to detect *Salmonella* at the species level. Herein, **the second aim of this thesis** was to develop and evaluate seven MCDA assays by targeting seven serovar/lineage-specific gene markers identified from the first aim. The developed MCDA assays would rapidly detect and differentiate the five most common *Salmonella* serovars that are prevalent in Australia and internationally.

*Shigella* and EIEC cause human bacillary dysentery with similar invasion mechanisms and share similar physiological, biochemical and genetic characteristics. The ability to differentiate *Shigella* and EIEC from each other is important for clinical diagnostics and public health epidemiologic investigations. The similarities between *Shigella* and EIEC strains make this differentiation very difficult as both share common ancestries within *E. coli*. However, *Shigella* and EIEC are phylogenetically separated into multiple clusters, making high resolution separation using cluster specific genomic markers possible. **The third aim of this thesis** was to identify phylogenetic clusters of *Shigella* and EIEC through large scale examination of publicly available genomes; and then identify cluster-specific gene markers using comparative genomic analysis of *Shigella* and EIEC

accessory genomes for differentiation of *Shigella* and EIEC and develop an automated pipeline for cluster typing and for *Shigella* and EIEC *in silico* serotyping based on the cluster-specific gene markers combined with *Shigella* and EIEC serotype-specific O antigen and H antigen genes from WGS data.

STEC infections have a significant impact on public health worldwide. STEC O157:H7 and “Big 6” non-O157:H7 serotypes are the major cause of foodborne outbreaks and human infections. Detection of STEC infections and determination of the serotype of the causing strain are important for accurate diagnosis and detection of outbreaks for public health control [213]. Previous phylogenetic analysis suggests that some STEC isolates form discrete clades associated with STEC sequence types and serotypes. Thus, **the fourth aim of this thesis** was to identify phylogenetic clusters of STEC through large scale examination of publicly available genomes; and identify cluster/serotype-specific genes for detection of STEC isolates and develop an automated pipeline for cluster typing and STEC *in silico* serotyping based on cluster/serotype-specific gene markers combined with *E. coli* O and H antigen genes from WGS data.

## Chapter 2. *In silico* Identification of Serovar-Specific Genes for *Salmonella* Serotyping

### 2.1 Link to thesis

In Chapter 1, I presented an overview of foodborne bacterial pathogens *Salmonella*, *Shigella* and STEC including their burden on the economy and human health, the existing identification and serotyping methods, the major challenge for detection and identification of these pathogens and the aims of this thesis. I also presented the shift towards identification of pathogen specific genomic markers for detection and serotyping of pathogens using genomics. For *Salmonella*, genes only found in the *Salmonella* serovar of interest have been identified previously but only for a small number of serovars. This prompted me to conduct genomic analysis on a large number of *Salmonella* genomes belonging to the 106 most common serovars as well as a number of rare serovars, for the purpose of identification of serovar-specific genes for these serovars. This chapter presents the first aim of this thesis.

I have published this work:

Zhang X, Payne M, Lan R. *In silico* Identification of Serovar-Specific Genes for *Salmonella* Serotyping. *Front Microbiol.* 2019;10:835.

I have presented this work at national conference:

Zhang X, Payne M, Lan R. *In silico* Identification of Serovar-Specific Genes for *Salmonella* Serotyping. Poster presentation, Australian Society for Microbiology Annual Scientific Meeting 2018.

**The Supplementary Material for this article can be found online at:**

<https://www.frontiersin.org/articles/10.3389/fmicb.2019.00835/full#supplementary-material>; or

[https://drive.google.com/drive/folders/1VkW2goYTdT\\_KYjlnEf4vCsnXCuW5iKt1?usp=sharing](https://drive.google.com/drive/folders/1VkW2goYTdT_KYjlnEf4vCsnXCuW5iKt1?usp=sharing).

## 2.2 Abstract

*Salmonella enterica* subspecies *enterica* is a highly diverse subspecies with more than 1500 serovars and the ability to distinguish serovars within this group is vital for surveillance. With the development of whole-genome sequencing technology, serovar prediction by traditional serotyping is being replaced by molecular serotyping. Existing *in silico* serovar prediction approaches utilize surface antigen encoding genes, core genome MLST and serovar-specific gene markers or DNA fragments for serotyping. However, these serovar-specific gene markers or DNA fragments only distinguished a small number of serovars. In this study, we compared 2258 *Salmonella* accessory genomes to identify 414 candidate serovar-specific or lineage-specific gene markers for 106 serovars which includes 24 polyphyletic serovars and the paraphyletic serovar Enteritidis. A combination of several lineage-specific gene markers can be used for the clear identification of the polyphyletic serovars and the paraphyletic serovar. We designed and evaluated an *in silico* serovar prediction approach by screening 1089 genomes representing 106 serovars against a set of 131 serovar-specific gene markers. The presence or absence of one or more serovar-specific gene markers was used to predict the serovar of an isolate from genomic data. We show that serovar-specific gene markers have comparable accuracy to other *in silico* serotyping methods with 84.8% of isolates assigned to the correct serovar with no false positives (FP) and false negatives (FN) and 10.5% of isolates assigned to a small subset of serovars containing the correct serovar with varied FP. Combined, 95.3% of genomes were correctly assigned to a serovar. This approach would be useful as diagnosis moves to culture-independent and metagenomic methods as well as providing a third alternative to confirm other genome-based analyses. The identification of a set of gene markers may also be useful in the development of more cost-effective molecular assays designed to detect specific gene markers of the all major serovars in a region. These assays would be useful in serotyping isolates where cultures are no longer obtained and traditional serotyping is therefore impossible.

**Keywords:** *Salmonella enterica*, accessory genomes, serotyping, serovar-specific gene markers, lineage-specific gene markers, polyphyletic serovars, paraphyletic serovar, serovar prediction

**Abbreviations:** FN, false negatives; FP, false positives; FPR, false positive rate; MLST, multi-locus sequence typing; NEPSS, National Enteric Pathogens Surveillance Scheme; PPV, positive predictive value; rSTs, ribosomal MLST STs; SISTR, *Salmonella in silico* typing resource; TN, true negatives; TNR, true negative rate; TP, true positives; TPR, true positive rate.

## 2.3 Introduction

*Salmonella* causes human salmonellosis and infections of warm-blooded animals (Kingsley and Bäumler, 2000). The *Salmonella* genus is divided into two species, *S. enterica* and *S. bongori*. serotyping further classifies *Salmonella* into over 2,600 serotypes (serovars) through the agglutination reaction of antisera to three surface antigens O, H1 and H2 (Le Minor and Bockemühl, 1984; Le Minor et al., 1990). There are 46 O antigens, that identify the serogroup. Together with 119 H1 and H2 flagellin antigens, the O, H1 and H2 combinations identify the serovars. Only a small proportion of the serovars are responsible for the majority of the human *Salmonella* infections (Popoff et al., 2004).

Serotyping by antigenic agglutination is being replaced by molecular serotyping (Cai et al., 2005; Wattiau et al., 2011). This can be achieved through examination of the sequence of O antigen gene cluster, H1 antigen encoding gene *fliC* and H2 antigen encoding gene *fliB* (Fitzgerald et al., 2007). O antigen gene clusters can be differentiated by presence or absence of genes while H1 and H2 antigens are differentiated by sequence variation (McQuiston et al., 2004; Guo et al., 2013; Zhang et al., 2015). *Salmonella* serotypes may also be inferred through multi-locus sequence typing (MLST) (Wattiau et al., 2011; Achtman et al., 2012) as a serotype may be inferred by its sequence types. However, a prerequisite for this approach is that prior knowledge of the corresponding relationship of serovar to sequence type is required.

Recently, with the development of whole-genome sequence-based comparison, several studies have identified genomic markers as an alternative molecular method for serotyping. Zou and colleagues (Zou et al., 2016) identified seven genes that provide

sufficient resolution to differentiate 309 *Salmonella* strains representing 26 serovars and found serovar-specific genes in 13 out of 26 serovars. Laing and colleagues (Laing et al., 2017) identified genomic fragments specific to *Salmonella* species and subspecies through pan-genome analysis. These specific genes or DNA fragments have been used as molecular targets to develop multiple molecular assays for rapid identification and detection of *Salmonella* at species and serovar level. However, these specific genes or DNA fragments are limited in their discriminative ability due to their ability to only distinguish a smaller number of serovars.

In this study, we aimed to use the extensive publicly available collection of *Salmonella* genomes to identify serovar-specific gene markers for the most frequent *Salmonella* serovars. We show the potential of these serovar-specific gene markers as markers for molecular serotyping either *in silico* typing of genomic data or for development of laboratory diagnostic methods.

## **2.4 Materials and methods**

### **2.4.1 Ribosomal MLST ST Based Isolate Selection**

The *Salmonella* database in the Enterobase (Alikhan et al., 2018) as of March 2018 was queried and 118997 isolate were examined. Representative isolates for each ribosomal MLST STs (rSTs) were selected and extracted by an in-house python script. Only serovars with more than 4 rSTs were included in this study. For the 20 largest serovars representative isolates were only randomly selected from rSTs with 2 or more isolates. For the remaining serovars, one representative isolate for each rST was randomly selected. Raw reads for these isolates were retrieved from ENA (European Nucleotide Archive, <https://www.ebi.ac.uk/ena>) and were *de novo* assembled using SPAdes v3.10.1 assembler with default settings [<http://bioinf.spbau.ru/spades>; (Bankevich et al., 2012)]. The serovar of the assembled genomes was predicted by *Salmonella in Silico* Typing Resource (SISTR) (Yoshida et al., 2016) after they met the following criteria which were defined by Robertson and colleagues (Robertson et al., 2018) using QUAST (<http://bioinf.spbau.ru/quast>) (Gurevich et al., 2013): assembly size between 4 and 6 Mb with the number of contigs less than 500, the largest contig greater than 100kb, GC content between 50% and 54%, gene predicted by glimmer within QUAST more than

3000. The concordance between the resulting SISTR serovar predictions and the reported serovar on the Enterobase metadata record were examined and a small number of genomes were removed from analysis due to inconsistent serovar predictions. The final data set consisted of 2258 high quality genomes with consistent serovar prediction representing 107 serovars (Supplementary Table S1).

#### **2.4.2 Identification of *Salmonella* Serovar-Specific Candidate Gene Markers**

To determine the potential serovar-specific gene markers for 107 serovars, the 2258 genomes were annotated using PROKKA (Seemann, 2014). Pan-genome and core-genome were analysed by roary (Page et al., 2015) using an 80% sequence identity threshold. The genes specific to each serovar were identified from the pan-genome's accessory genes with an in-house python script. In this study, the number of genomes from a given serovar containing a specific gene for that serovar was termed true positive (TP), the number of genomes from the same serovar lacking the same gene was termed false negative (FN). The number of genomes from other serovars containing the same serovar-specific gene was termed false positive (FP). Relaxed cutoffs (20% FN, 10% FP) were used initially in order to ensure that all serovars had candidate specific genes which could be further investigated. Paralogous genes were removed from the analyses.

#### **2.4.3 Evaluation of Potential Serovar-Specific Gene Markers**

The  $F_1$  score was used for initial selection of the potential serovar-specific gene markers.  $F_1$  score was evaluated based on the formula:  $2 \times (\text{PPV} \times \text{Sensitivity}) / (\text{PPV} + \text{Sensitivity})$ , where PPV standing for positive predictive value which was defined as  $\text{TP}/(\text{TP}+\text{FP})$  and sensitivity [true positive rate (TPR)] was defined as  $\text{TP}/(\text{TP}+\text{FN})$ . The  $F_1$  ranges from 0 to 1, where 1 means the serovar-specific gene which was present in all genomes of a given serovar and absent in all genomes of other serovars. The serovar-specific gene markers were selected using the best performing gene for each serovar based on  $F_1$  score. The specificity [True negative rate (TNR)] defined as  $\text{TN}/(\text{TN}+\text{FP})$  was used to evaluate true negative (TN) rate of serovar-specific gene markers. False positive rate (FPR) was defined by  $1 - \text{TNR}$ .



#### **2.4.4 Phylogenetic Analyses**

In order to determine the causes for the observed false negative and false positive rates in the candidate serovar-specific gene markers, the phylogenetic relationships of the serovars involved were investigated. The draft assemblies of 1258 isolates were used to generate phylogenetic trees by using ParSNP v1.2 (<http://github.com/marbl/harvest>) (Treangen et al., 2014) with default parameters to determine the phylogeny between and within serovars. The tree was visualised by Figtree v1.4.3 (Schneider et al., 2000).

#### **2.4.5 Location and Functions of Serovar-Specific Gene Markers**

Representative complete genomes for each serovar containing gene features were downloaded from NCBI (<https://www.ncbi.nlm.nih.gov/>) and were used to determine the location of each of candidate serovar-specific gene by BLASTN with default settings (version 2.2.6, Supplementary Table S2). In serovars with no representative complete genome a representative genome was selected from isolates assembled in this study. Sequences of serovar-specific gene markers are included in Supplementary Data S1. Clustering of genes across the genome was used to investigate whether the serovar-specific gene markers were potentially part of a single element gained by a serovar in one event. The candidate serovar-specific gene markers were considered as a cluster if they were located less than 5kb from each other.

The functional categories of gene markers were identified from RAST annotation (<http://rast.theseed.org/FIG/rast.cgi>) (Aziz et al., 2008). The prophage sequences within serovars reference genomes were identified by using PHASTER to indicate whether the serovar-specific gene markers may have been acquired along with prophages (PHAge Search Tool Enhanced Release) (Arndt et al., 2016).

#### **2.4.6 *In silico* Serotype Prediction Using Serovar-Specific Gene Markers**

An additional 1089 isolates were selected from the Enterobase using an in-house python script with the exclusion of 2258 isolates used for the initial screening from the same database as of March 2018 (Supplementary Table S3). BLASTN was used to search against the 1089 genomes belonging to 106 *Salmonella* serovars for the presence of any of the serovar-specific gene markers. Custom python scripts were then used to predict serovar from these serovar assignments based on the known gene presence pattern for

each serovar. The TP was classified as the total number of correctly assigned serovars and cases where the correct serovar was called as well as one or more false positives. Failed assignment was defined where no serovar or incorrect serovars were called. Serovar predictions were compared to SeqSero (Zhang et al., 2015) and SISTR predictions.

#### **2.4.7 Calculation of the Specificity of Candidate Serovar-Specific Gene Markers for Common Serovars**

The specificity of typing rate for common serovars (Hendriksen et al., 2011) was equal to  $(1 - \text{potential error rate})$ . The potential error rate of serovar-specific gene markers defined by the formula:  $(\text{Number of FPs}) * (\text{The frequency of that serovar in a given region}) / (\text{Total of genomes of that serovar})$ .

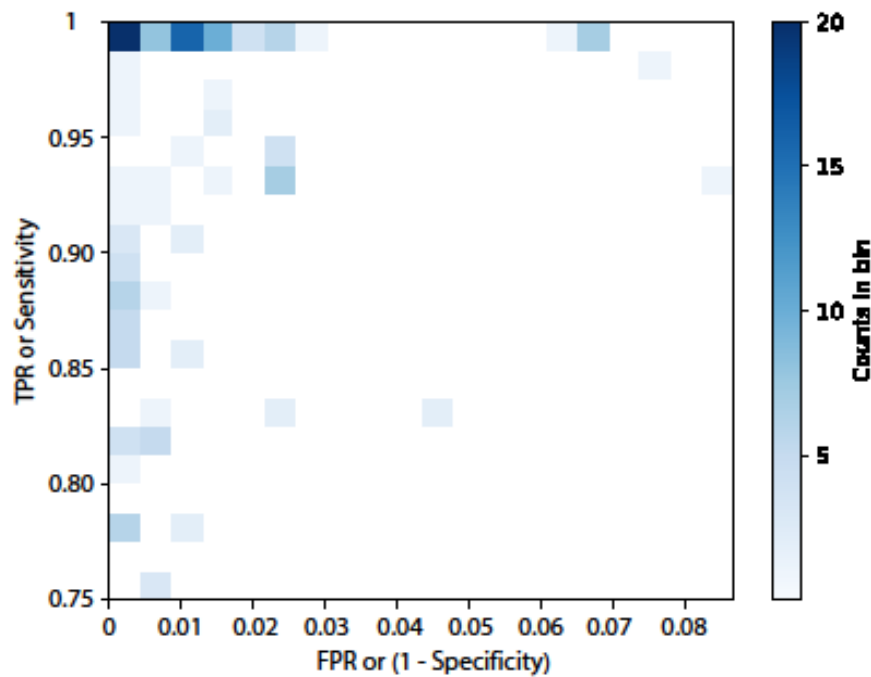
## **2.5 Results**

### **2.5.1 Identification of Candidate Serovar-Specific Gene Markers**

The accessory genes from 2258 genomes representing 107 serovars were screened to identify potential serovar-specific gene markers. This initial screening identified 354 potential serovar-specific gene markers within 101 serovars. Six serovars namely, Bareilly, Bovismorbificans, Thompson, Reading, Typhi, and Saintpaul had no candidate serovar-specific gene markers that were present in all lineages of a given serovar. The specificity (True negative rate) and sensitivity (True positive rate) of the 354 candidate serovar-specific gene markers were also examined and summarised in Figure 1. Forty serovars contained 194 serovar-specific gene markers with 100% specificity and sensitivity (no FN or FP), while 31 serovars contained 80 candidate serovar-specific gene markers with 100% sensitivity but with less than 100% specificity (varied FP). Nine serovars contained 27 candidate serovar-specific gene markers with 100% specificity but with less than 100% sensitivity (varied FN). The remaining 21 serovars contained 53 candidate serovar-specific gene markers with both specificity and sensitivity less than 100% (varied FN and FP).

We constructed a phylogenetic tree using 1258 representative isolates from 107 serovars using ParSNP (Supplementary Figure S1). The 1258 isolates were selected based on

phylogenetic relationships of the initial 2258 isolates from which we selected isolates to represent each independent lineage. We found that members of each of the 82 serovars formed a monophyletic lineage while 24 serovars were polyphyletic with each made up of 2 to 4 lineages. Several of these serovars are known to be polyphyletic and are unlikely to contain serovar-specific gene markers (Falush et al., 2006; den Bakker et al., 2011; Timme et al., 2013; Graham et al., 2018). Serovar Enteritidis is paraphyletic with three other serovars (Dublin, Berta and Gallinarium) arising from within the larger Enteritidis clade which is itself made up of three lineages known as clade A, B and C (Wattiau et al., 2008). The five Enteritidis-specific candidate gene markers were negative to the Enteritidis isolates which clustered separately on the tree.



**Chapter 2. Figure 1: The distribution of sensitivity and specificity of 354 potential serovar-specific gene markers.** TPR, true positive rate. FPR, false positive rate. Where a gradient from light blue (low percentage) to dark blue (high percentage) is displayed.

Interestingly for four polyphyletic serovars, Bredeney, Kottbus, Livingstone and Virchow, each had one candidate serovar-specific gene which was present in all isolates of that serovar. For the remaining 20 polyphyletic serovars and paraphyletic serovar Enteritidis, we searched for lineage-specific gene markers as each serovar contained more than one lineage. If all lineages contained at least one lineage-specific gene, we regard

that serovar as containing serovar-specific gene markers. A total of 111 potential lineage-specific gene markers were identified for 19 polyphyletic serovars and paraphyletic serovar Enteritidis, among which, 27 lineage-specific gene markers were identified for 5 serovars with 100% specificity and sensitivity (no FN and FP), 76 candidate lineage-specific gene markers for 14 serovars with 100% sensitivity and less than 100% specificity (varied FP), and Enteritidis containing 6 candidate lineage-specific gene markers with varied FN and FP (Table 1).

For the 11 of the 82 monophyletic serovars that lacked serovar-specific candidate gene markers due to false negatives, we found that the FN was often due to isolates that are grouped on one branch and diverged earlier from the other isolates. For such groups, we searched for lineage-specific gene markers. Therefore, two or more gene markers can be used to identify a serovar and such serovars were also considered to contain serovar-specific gene markers, similar to polyphyletic serovars. Three serovars, Paratyphi A, Heidelberg and Muenchen could be identified by the combined lineage-specific gene markers.

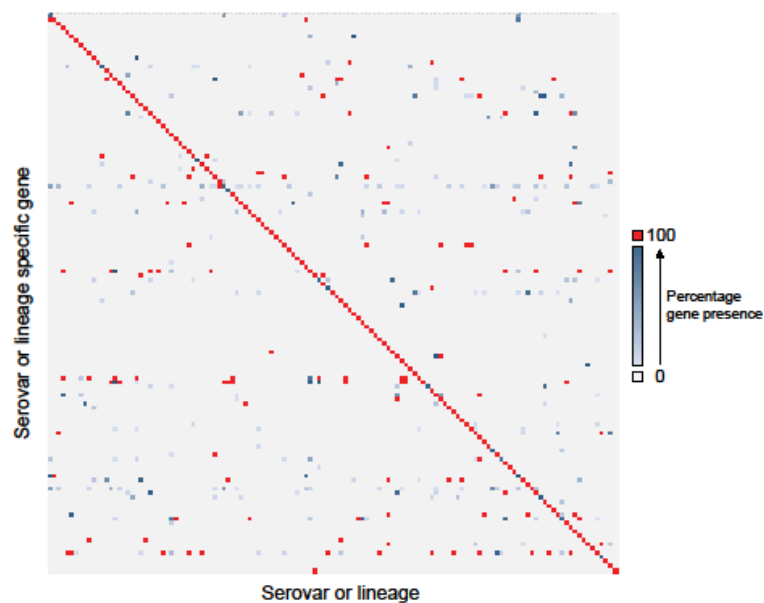
A total of 414 candidate serovar-specific gene markers including 295 serovar-specific gene markers and 119 lineage-specific gene markers are summarised in Supplementary Table S2. In total, 106 of 107 serovars contained 1 or more gene markers, 33 serovars contained one specific gene while 73 contained two or more gene markers. There were no candidate serovar-specific gene markers found for monophyletic Typhi and no potential lineage-specific gene markers found for lineage III of Stanleyville which contained only one isolate.

### **2.5.2 Functional Categories of Serovar-Specific Gene Markers**

Functional characterisation of all 414 gene markers identified for the 106 serovars using RAST found that 197 had known functions and 217 encoded hypothetical proteins with unknown functions. Only 46 genes with annotations can be grouped into functional categories while 151 genes with functions were not in RAST functional categories (Table 2). Using PHASTER, 45 candidate serovar-specific gene markers were located within predicted prophages.

### 2.5.3 A Minimal Set of Serovar-Specific Gene Markers for *in silico* Molecular Serotyping

For many serovars, multiple candidate serovar-specific gene markers or lineage-specific gene markers were identified. In these cases, a single gene was selected that has the lowest FN and FP rates. A minimum of 131 gene markers allows identification of the serovars with error rates from 0 to 8.33%. The distribution of the gene markers across all 106 serovars demonstrates high degree of specificity as shown in Figure 2 in which the diagonal displays the one to one relationship of the serovar or lineage with serovar-specific gene markers while the off-diagonal space showed sparse scattered presence of



**Chapter 2. Figure 2: The distribution of a minimal set of 131 serovar-specific genes in 106 serovars.** The Y-axis shows serovar or lineage-specific gene markers and the X-axis shows serovars or lineages. The details were listed in supplementary table 4. Grey indicated zero genomes containing a gene (true negatives). Gene/Genome pairs along the diagonal represent genomes containing the serovar-specific gene markers that matches their serovar (True positives). Red represents genes that are present in 100% of genomes for a given serovar or lineage. Where a gene is present in less than 100% of a serovar a gradient from light blue (low percentage) to dark blue (high percentage) is displayed. Blue pairs along the diagonal represent the presence of false negatives. Pairs that are blue or red outside of the diagonal represent pairs containing genes that do not match the predicted serovar of the genome (false positives).

**Chapter 2. Table 1: Lineage-specific candidate gene markers for polyphyletic serovars and paraphyletic serovar**

<b>Serovar</b>	<b>No of genomes</b>	<b>No of lineage</b>	<b>Lineages</b>	<b>No of genes</b>	<b>Sensitivity<sup>#</sup></b>	<b>Specificity<sup>#</sup></b>
Bareilly	20	2	Bareilly-I	2	100	98.76
			Bareilly-II	1	100	99.11
Bovismorbificans	34	2	Bovismorbificans-I	1	100	97.25
			Bovismorbificans-II	1	100	99.91
Bredeney	5	2	Bredeney	1	100	97.61
Cerro	40	2	Cerro-I	4	100	100
			Cerro-II	2	100	100
Derby	24	3	Derby-I&II	1	100	100
			Derby-III	4	100	100
Enteritidis	165	2	Enteritidis-clade A/C	1	100	98.85
			Enteritidis-clade B	5	96.43*	99.65
Give	26	3	Give-I&II	4	100	94.6
			Give-III	1	100	99.82
Havana	20	2	Havana-I	2	100	97.39
			Havana-II	4	100	100
Hvittingfoss	16	3	Hvittingfoss-I&II	1	100	100
			Hvittingfoss-III	1	100	100
Kentucky	31	2	Kentucky-I	5	100	100
			Kentucky-II	3	100	100
Kottbus	12	3	Kottbus	1	100	93.98
Livingstone	17	2	Livingstone	1	88.24*	99.47
London	11	2	London-I	2	100	99.11

			London-II	3	100	99.87
Mississippi	14	2	Mississippi-I	5	100	100
			Mississippi-II	1	100	100
Newport	85	3	Newport-I&II	1	100	92.87
			Newport-I&III	1	100	91.67
Oranienburg	29	4	Oranienburg-I&II&IV	1	100	98.67
			Oranienburg-III	1	100	98.72
Oslo	9	2	Oslo-I	2	100	99.91
			Oslo-II	1	100	100
Paratyphi B	72	3	Paratyphi B-I&II	11	100	97.83
			Paratyphi B-III	1	100	100
			Paratyphi B-mono	1	100	100
Reading	8	2	Reading-I	1	100	100
			Reading-II	2	100	99.96
Saintpaul	31	3	Saintpaul-I	11	100	98.14
			Saintpaul-II	5	100	100
			Saintpaul-III	1	100	98.27
Senftenberg	27	3	Senftenberg-I&II	2	100	99.96
			Senftenberg-III	1	100	100
Stanleyville	6	3	Stanleyville-I&II	2	83.33*	95.44
Telelkebir	8	2	Telelkebir-I	3	100	100
			Telelkebir-II	6	100	100
Thompson	32	2	Thompson-I	2	100	98.49
			Thompson-II	2	100	100
Virchow	39	2	Virchow	1	100	100

\*: The sensitivity of less than 100% was due to at least one target serovar genome lacking the candidate gene. Six out of 165 isolates of Enteritidis, two out of 17 isolates of Livingstone-I and one out of 6 isolates of Stanleyville-III lacked candidate lineage-specific gene markers.

#: Sensitivity and specificity for the best performing gene for each lineage. The number of isolates used to arrive at Sensitivity and Specificity calculation for each serovar-specific gene marker were listed in supplementary Table S2.

**Chapter 2. Table 2: Serovar-specific genes functional categories**

Category by RAST	No of genes*
DNA Metabolism	18
Regulation and cell signalling	5
Carbohydrates	2
Membrane Transport	8
Virulence, Disease and Defence	1
RNA Metabolism	4
Stress Response	2
Cofactors, Vitamins, Prosthetic Groups, Pigments	1
Cell Wall and Capsule	1
Phages related	2
Protein Metabolism	1
Amino Acids and Derivatives	1
Uncategorized	152
Hypothetical proteins with unknown function	217

\*: The details of these genes were listed in Table S2.



these genes in other serovars of varied percentages indicating a low false positive rate. The details of these gene markers were listed in Supplementary Table S4. Overall, 45 serovars can be distinguished by their respective serovar-specific gene and 61 serovars can be differentiated by a combination of gene markers.

We tested an additional 1089 genomes belonging to 106 non-typhoidal *Salmonella* serovars to evaluate the ability of the 131 specific gene markers to correctly assign serovars to isolates. Using the serovar-specific gene markers, 1038 of the 1089 isolates (95.3%) were successfully assigned (924 to correct serovar with no false positives or false negatives [84.8%] and 114 to the correct serovar with some false positives [10.5%]) and 51 failed (4.7%). For SISTR and SeqSero, the number of concordant serovar assignments were 1037 (95%) and 905 (82.8%) respectively (Supplementary Table S3).

#### **2.5.4 Serovar-Specific Gene Markers for Serotyping of Common Serovars**

The top 20 serovars causing human infection found in each continent (Hendriksen et al., 2011) were collapsed into a combined list of 46 serovars (Supplementary Table S5). Since these serovars contained the vast majority of isolates causing human infections globally, we consider them separately to assess the utility of candidate serovar-specific gene markers for serotyping of most prevalent serovars in a local setting. When only these serovars were considered, 18 out of 46 could be uniquely identified by one of the serovar-specific gene markers. To increase accuracy of typing in the remaining 28 common serovars where serovar-specific gene markers have varied false positive rates, we examined using subsets of the 131 gene markers (ranging from 2 to 9 genes per serovar) to eliminate potential false positives. For example, the combination of Choleraesuis specific gene and Cerro-I lineage-specific gene can eliminate false positive isolate of Cerro from Choleraesuis, if both genes are positive, the isolate could be assigned Cerro while if Cerro-I lineage-specific gene is negative, the isolate is Choleraesuis.

To estimate potential errors in typing, we took into account the frequency of the 46 common serovars that showed large differences between regions (Hendriksen et al., 2011). Therefore, different combinations of genes may be used to specifically limit false positive results from serovars present in that region. In a given region, the specificity of common candidate serovar-specific gene markers was calculated using the rate of false positives

and the frequency of the false positive serovar in that region. The specificity of candidate serovar-specific gene markers was also calculated using the FP rate (Supplementary Table 4). For example, a panel of 15 genes could be used for typing the 10 most frequent serovars in Australia (NEPSS 2010) (Table 3). When Australian regional frequencies were taken into account, the genes listed in Table 3 can be used as markers for laboratory based typing and the error rate will be less than 2.4%.

## 2.6 Discussion

*Salmonella* serotyping has been vital for diagnosis and surveillance. Serovar prediction by traditional serotyping can be limited by the lack of surface antigen expression or autoagglutination properties (Wattiau et al., 2008). Recently, with the development of whole-genome sequencing technology, the relevant genomic regions of the *rfb* gene cluster for O antigen, gene *fliC* and gene *fliB* for H antigens, and genes targeted by MLST can be extracted and used for serovar identification. Several studies have identified serovar-specific genes or DNA fragments for serotyping through whole-genome sequencing based genomic comparison (Zou et al., 2013; Zou et al., 2016; Laing et al., 2017). However, these serovar-specific genes or DNA fragments only distinguished a smaller number of serovars. In this study, we identified 414 candidate serovar-specific or lineage-specific gene markers for 106 serovars which includes 24 polyphyletic serovars and the paraphyletic serovar Enteritidis. A subset of these gene markers were validated by independent genomes and were able to assign serovars correctly in 95.3% of cases.

The above analysis was complicated by the presence of polyphyletic serovars, which arise independently from separate ancestors to form separate lineages. Therefore, a combination of lineage-specific gene markers was required for the clear identification of the majority of the polyphyletic serovars. Interestingly four polyphyletic serovars, Bredeney, Kottbus, Livingstone and Virchow, each had one candidate serovar-specific gene marker which was present in all isolates of that serovar. The Bredeney serovar-specific gene was predicted to encode a translocase involved in O-antigen conversion and could have been gained in parallel. The serovar-specific genes of the other three polyphyletic serovars encode hypothetical proteins with unknown function and no apparent explanation for their presence in different lineages of the same serovar.

**Chapter 2. Table 3: A panel of serovar-specific genes for typing the ten most frequent serovars in Australia**

<b>Serovar</b>	<b>Gene 1</b>	<b>Gene 2</b>	<b>Gene 3</b>	<b>Gene 4</b>	<b>Gene 5</b>	<b>Gene 6</b>	<b>Gene 7</b>	<b>Gene 8</b>	<b>Gene 9</b>	<b>Gene 10</b>	<b>Gene 11</b>	<b>Gene 12</b>	<b>Gene 13</b>	<b>Gene 14</b>	<b>Gene 15</b>
Typhimurium	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Enteritidis-B	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-
Enteritidis-A/C	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-
Virchow	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-
Saintpaul-I	-	-	-	-	+	-	-	-	[+]	-	-	-	-	-	-
Saintpaul-II	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-
Saintpaul-III	[+]	-	-	-	-	-	+	-	-	-	-	-	-	-	-
Infantis	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-
Paratyphi B-I&II	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-
Paratyphi B-III	[+]	-	-	-	-	-	-	-	-	+	-	-	-	-	-
Chester	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-
Hvittingfoss-I&II	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-
Hvittingfoss-III	[+]	-	-	-	-	-	-	-	-	-	-	-	+	-	-
Muenchen-I	-	-	-	-	-	-	[+]	-	-	-	-	-	-	+	-
Muenchen-II	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+
Error rate	2.4	0	1.5	0	2.9	0	0.2	0	1	0	2.2	0	0	0	0.9
Specificity	97.6	100	98.5	100	97.1	100	99.8	100	99	100	97.8	100	100	100	99.1

"+": true positives (TP); "-": true negatives (TN); [+]: false positives (FP) in a subset of genomes. Gene 1 = STM4494 (Typhimurium); Gene 2 = SEN1384 (Enteritidis-clade B); Gene 3 = R561\_RS18155 (Enteritidis-clade A/C); Gene 4 = SEV\_RS01820 (Virchow); Gene 5 = SESP\_A\_RS08460 (Saintpaul-I); Gene 6 = SeSPB\_A1749 (Saintpaul-II); Gene 7 = Saintpaul-III; Gene 8 = L287\_RS37190 (Infantis); Gene 9 = SPAB\_01124 (Paratyphi B-I&II); Gene 10 = SPAB\_01338 (Paratyphi B-III); Gene 11 = SEECH997\_RS20295 (Chester); Gene 12 = LFZ15\_01345 (Hvittingfoss-I&II); Gene 13 = LFZ15\_20305 (Hvittingfoss-III); Gene 14 = L098\_RS21065 (Muenchen-I); Gene 15 = Muenchen-II. See Supplementary Table S2 for gene details. The potential error rate of serovar-specific genes was defined by the formula:  $(\text{Number of FPs}) * (\text{The frequency of that serovar in a given region}) / (\text{Total of genomes of that serovar})$ . The specificity of typing rate was equal to  $(1 - \text{potential error rate})$ .

Unlike polyphyletic serovars, the three lineages (clade A, B and C) of the paraphyletic serovar Enteritidis share a recent common ancestor. Clade A and C are ancestral to Clade B. Previous studies described that Enteritidis was clustered with serovars Dublin, Berta and Gallinarium which was called “Section Enteritidis” (Vernikos et al., 2007; Achtman et al., 2012; Allard et al., 2013; Timme et al., 2013). Another study showed that serovar Nitra was embedded within Enteritidis lineages by using whole genome phylogeny (Deng et al., 2014). There also was cross-reactivity between Enteritidis and Nitra according to Ogunremi’s study (Ogunremi et al., 2017). In our study, we selected the isolates based on rSTs, Nitra was not present in Enterobase rMLST database when this study commenced and so was not included in this study. Gallinarium is distinguishable from Enteritidis using the presence of a 4 bp deletion in the *speC* gene (Kang et al., 2011). We observed that the common ancestors of serovars Dublin, Berta and Gallinarium, arose from an ancestor between Clades B and A/C. While Dublin can be separately identified, we cannot distinguish Berta or Gallinarium from Enteritidis clade A/C. These results highlight a limitation of the approach as serovars must be sufficiently divergent that they differ by at least one unique gene. Similarly, there were 8 other serovars that were not distinguishable likely due to very recent shared ancestry with little gene acquisition.

Serovar-specific candidate gene markers or lineage-specific candidate gene markers in 69 out of 106 serovars were contiguous in the genome with similar functions grouped together (data not shown). This suggests that these gene markers may have been incorporated into serovar genomes together through horizontal gene transfer. Indeed the seven Typhimurium specific candidate gene markers identified in this study (STM4492, STM4493, STM4494, STM4495, STM4496, STM4497, STM4498) were located in Typhimurium tRNA<sup>leuX</sup> integrating conjugative element-related region including genes from STM4488 to STM 4498, which is a known horizontal gene transfer hotspot (Bishop et al., 2005). Similarly five Enteritidis specific candidate gene markers identified (SEN1379, SEN1380, SEN1382, SEN1383, SEN1383) were located in the Sdr I region (Agron et al., 2001) and the prophage-like GEI/φSE14 region (Santiviago et al., 2010). Both of these regions are linked to prophages, which suggests that these regions integrated into the genome of a common ancestor of the global Enteritidis clade and were derived from horizontal gene transfer.

Other methods for *in silico* serovar prediction are implemented in SeqSero (Zhang et al., 2015) and SISTR (Yoshida et al., 2016). Both of these methods examine genomic regions responsible for surface antigens while SISTR also implements a cgMLST scheme to examine overall genetic relatedness. Additionally, traditional 7 gene MLST and eBURST groups derived from it can also be used for *in silico* serovar determination (Achtman et al., 2012; Ashton et al., 2016; Robertson et al., 2018). Both SISTR and SeqSero provide higher discriminatory power than traditional serovar identification (Yachison et al., 2017). However, they have a number of drawbacks such as indistinguishable serovars having the same antigenic formula or antigenic determinants not being expressed (Robertson et al., 2018). In the current study, we examined *in silico* serovar prediction by screening genomes against a set of 131 serovar-specific gene markers. The approach provided serovar prediction by yielding “presence or absence” of individual serovar-specific gene marker or combination of gene markers in a query isolate. We show that serovar-specific gene markers have comparable accuracy to other *in silico* serotyping methods with 91.5% isolates from initial identification dataset and 84.8% isolates from a validation dataset assigned to the correct serovar (with no FN and FP). 10.5% of isolates from validation dataset can be assigned to a small subset of serovars containing the correct serovar (with varied FP). The specificity for *in silico* serovar prediction approach by serovar-specific gene markers was 95.3%, slightly higher than SISTR (95%) and SeqSero (82.8%) in the same dataset we tested. This result was similar to the specificities of SISTR and SeqSero reported by Yachison and colleagues which were 94.8% and 88.2% respectively (Yachison et al., 2017).

Our serovar-specific gene marker based method does not require the accurate examination of O antigen gene clusters or sequence variation of the H antigen genes which can be problematic. Our method also alleviates the need for the entire gene or genome sequence be assembled which is necessary in MLST or cgMLST based methods. Therefore, this approach may be useful for cases where very little sequence is available such as in metagenomics or culture free typing as well as providing a third alternative to confirm other analyses.

The identification of a set of gene markers able to uniquely identify all prevalent serovars in a region may also be useful in the development molecular assays. These assays would be useful in serotyping isolates where cultures are no longer obtained and traditional serotyping is therefore impossible. For example, a set of PCR assays could be designed that would allow the sensitive detection of specific gene markers, and therefore allow prediction of the serovar, from a clinical sample. Additionally, by eliminating the need to detect serovars that are very rarely observed in a region the number of these gene markers required to detect all major serovars in a region can be significantly reduced allowing for a more cost-effective assay.

## **2.7 Conclusion**

In this study we identified candidate serovar-specific gene markers and candidate lineage-specific gene markers for 106 serovars by characterising the accessory genomes of a representative selection of 2258 strains as potential markers for *in silico* serotyping. We account for polyphyletic and paraphyletic serovars to provide a new method, using the presence or absence of these gene markers, to predict the serovar of an isolate from genomic data. The gene markers identified here may also be used to develop serotyping assays in the absence of an isolated strain which will be useful as diagnosis moves to culture independent and metagenomic methods.

## **2.8 Author contributions**

MP and RL designed the study. XZ and MP performed the bioinformatic analysis. XZ, MP and RL analysed the results. XZ drafted the manuscript. MP and RL provided critical revision of the manuscript.

## **2.9 Funding**

This work was supported a National Health and Medical Research Council project grant.

## 2.10 Supplementary material

The Supplementary Material for this article can be found online at:

<https://www.frontiersin.org/articles/10.3389/fmicb.2019.00835/full#supplementary-material>

**FIGURE S1** | The SNP based phylogenetic tree constructed by ParSNP showing the evolutionary relationships within and between serovars using 1344 representative isolates including 1258 isolates from 107 serovars examined in the study and 86 isolates from serovars with less than 5 rSTs which were otherwise excluded from the study.

**TABLE S1** | The final data set of 2258 high quality and consistent serovar prediction genomes representing 107 serovars.

**TABLE S2** | A total of 414 candidate serovar-specific genes including 295 serovar-specific genes and 119 lineage-specific genes.

**TABLE S3** | An additional 1089 validation isolates with serovar prediction results by SISTR, SeqSero and serovar-specific gene markers.

**TABLE S4** | A minimum of 131 genes for identification of 106 serovars.

**TABLE S5** | A set of 65 genes for identification of 46 common serovars.

**DATA S1** | Sequences of 131 serovar-specific gene markers.

## 2.11 References

- Achtman, M., Wain, J., Weill, F.X., Nair, S., Zhou, Z., Sangal, V., et al. (2012). Multilocus sequence typing as a replacement for serotyping in *Salmonella enterica*. *PLoS Pathog.* 8:e1002776. doi: 10.1371/journal.ppat.1002776
- Agron, P.G., Walker, R.L., Kinde, H., Sawyer, S.J., Hayes, D.C., Wollard, J., et al. (2001). Identification by subtractive hybridization of sequences specific for *Salmonella enterica* serovar enteritidis. *Appl. Environ. Microbiol.* 67, 4984-4991. doi: 10.1128/AME.67.11.4984-4991.2001
- Alikhan, N.F., Zhou, Z., Sergeant, M.J., and Achtman, M. (2018). A genomic overview of the population structure of *Salmonella*. *PLoS Genet.* 14:e1007261. doi: 10.1371/journal.pgen.1007261
- Allard, M.W., Luo, Y., Strain, E., Pettengill, J., Timme, R., Wang, C., et al. (2013). On the evolutionary history, population genetics and diversity among isolates of



- Salmonella* Enteritidis PFGE pattern JEGX01.0004. *PLoS One* 8:e55254. doi: 10.1371/journal.pone.0055254
- Arndt, D., Grant, J.R., Marcu, A., Sajed, T., Pon, A., Liang, Y., et al. (2016). PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic Acids Res.* 44, W16-W21. doi: 10.1093/nar/gkw387
- Ashton, P.M., Nair, S., Peters, T.M., Bale, J.A., Powell, D.G., Painset, A., et al. (2016). Identification of *Salmonella* for public health surveillance using whole genome sequencing. *PeerJ* 4:e1752. doi: 10.7717/peerj.1752
- Aziz, R.K., Bartels, D., Best, A.A., DeJongh, M., Disz, T., Edwards, R.A., et al. (2008). The RAST Server: rapid annotations using subsystems technology. *BMC Genomics* 9:75. doi: 10.1186/1471-2164-9-75
- Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., et al. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol.* 19, 455-477. doi: 10.1089/cmb.2012.0021
- Bishop, A.L., Baker, S., Jenks, S., Fookes, M., Gaora, P.O., Pickard, D., et al. (2005). Analysis of the hypervariable region of the *Salmonella enterica* genome associated with tRNA<sup>LeuX</sup>. *J. Bacteriol.* 187, 2469-2482. doi: 10.1128/JB.187.7.2469-2482.2005
- Cai, H., Lu, L., Muckle, C., Prescott, J., and Chen, S. (2005). Development of a novel protein microarray method for serotyping *Salmonella enterica* strains. *J. Clin. Microbiol.* 43, 3427-3430. doi: 10.1128/JCM.43.7.3427-3430.2005
- den Bakker, H.C., Moreno Switt, A.I., Govoni, G., Cummings, C.A., Ranieri, M.L., Degoricija, L., et al. (2011). Genome sequencing reveals diversification of virulence factor content and possible host adaptation in distinct subpopulations of *Salmonella enterica*. *BMC Genomics* 12:425. doi: 10.1186/1471-2164-12-425
- Deng, X., Desai, P.T., den Bakker, H.C., Mikoleit, M., Tolar, B., Trees, E., et al. (2014). Genomic epidemiology of *Salmonella enterica* serotype Enteritidis based on population structure of prevalent lineages. *Emerg. Infect. Dis.* 20, 1481-1489. doi: 10.3201/eid2009.131095
- Falush, D., Torpdahl, M., Didelot, X., Conrad, D.F., Wilson, D.J., and Achtman, M. (2006). Mismatch induced speciation in *Salmonella*: model and data. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 361, 2045-2053. doi: 10.1098/rstb.2006.1925

- Fitzgerald, C., Collins, M., van Duyn, S., Mikoleit, M., Brown, T., and Fields, P. (2007). Multiplex, bead-based suspension array for molecular determination of common *Salmonella* serogroups. *J. Clin. Microbiol.* 45, 3323-3334. doi: 10.1128/JCM.00025-07
- Graham, R.M.A., Hiley, L., Rathnayake, I.U., and Jennison, A.V. (2018). Comparative genomics identifies distinct lineages of *S. Enteritidis* from Queensland, Australia. *PLoS One* 13:e0191042. doi: 10.1371/journal.pone.0191042
- Guo, D., Liu, B., Liu, F., Cao, B., Chen, M., Hao, X., et al. (2013). Development of a DNA microarray for molecular identification of all 46 *Salmonella* O serogroups. *AEM* 79, 3392-3399. doi: 10.1128/AEM.00225-13
- Gurevich, A., Saveliev, V., Vyahhi, N., and Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. *Bioinformatics* 29, 1072-1075. doi: 10.1093/bioinformatics/btt086
- Hendriksen, R.S., Vieira, A.R., Karlsmose, S., Lo Fo Wong, D.M., Jensen, A.B., Wegener, H.C., et al. (2011). Global monitoring of *Salmonella* serovar distribution from the World Health Organization Global Foodborne Infections Network Country Data Bank: results of quality assured laboratories from 2001 to 2007. *Foodborne Pathog. Dis.* 8, 887-900. doi: 10.1089/fpd.2010.0787
- Kang, M.S., Kwon, Y.K., Jung, B.Y., Kim, A., Lee, K.M., An, B.K., et al. (2011). Differential identification of *Salmonella enterica* subsp. *enterica* serovar Gallinarum biovars Gallinarum and Pullorum based on polymorphic regions of *glgC* and *speC* genes. *Vet Microbiol.* 147, 181-185. doi: 10.1016/j.vetmic.2010.05.039
- Kingsley, R.A., and Bäumler, A.J. (2000). Host adaptation and the emergence of infectious disease: the *Salmonella* paradigm. *Mol Microbiol.* 36, 1006-1014. doi: 10.1046/j.1365-2958.2000.01907.x
- Laing, C.R., Whiteside, M.D., and Gannon, V.P.J. (2017). Pan-genome Analyses of the Species *Salmonella enterica*, and Identification of Genomic Markers Predictive for Species, Subspecies, and Serovar. *Front. Microbiol* 8, 1345. doi: 10.3389/fmicb.2017.01345
- Le Minor, L., and Bockemühl, J. (1984). Supplement No XXVII au schéma de Kauffmann-White. *Ann. Institut Pasteur Microbiol.* 135b, 45-51. doi: 10.1016/S0769-2609(84)80042-3

- Le Minor, L., Popoff, M., and Bockemühl, J. (1990). Supplement 1989 (n° 33) to the Kauffmann-White scheme. *Res. Microbiol.* 141, 1173–1177. doi: 10.1016/0923-2508(90)90090-D
- McQuiston, J.R., Parrenas, R., Ortiz-Rivera, M., Gheesling, L., Brenner, F., and Fields, P.I. (2004). Sequencing and comparative analysis of flagellin genes *fliC*, *fljB*, and *flpA* from *Salmonella*. *J. Clin. Microbiol.* 42, 1923-1932. doi: 10.1128/JCM.42.5.1923-1932.2004
- Ogunremi, D., Nadin-Davis, S., Dupras, A.A., Márquez, I.G., Omid, K., Pope, L., et al. (2017). Evaluation of a Multiplex PCR Assay for the Identification of *Salmonella* Serovars Enteritidis and Typhimurium Using Retail and Abattoir Samples. *J. Food Prot.* 80, 295-301. doi: 10.4315/0362-028X.JFP-16-167.
- Page, A.J., Cummins, C.A., Hunt, M., Wong, V.K., Reuter, S., Holden, M.T., et al. (2015). Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* 31, 3691-3693. doi: 10.1093/bioinformatics/btv421
- Popoff, M.Y., Bockemühl, J., and Gheesling, L.L. (2004). Supplement 2002 (no. 46) to the Kauffmann-White scheme. *Res. Microbiol.* 155, 568-570. doi: 10.1016/j.resmic.2004.04.005
- Robertson, J., Yoshida, C., Kruczkiewicz, P., Nadon, C., Nichani, A., Taboada, E.N., et al. (2018). Comprehensive assessment of the quality of *Salmonella* whole genome sequence data available in public sequence databases using the *Salmonella in silico* Typing Resource (SISTR). *Microb. Genomics* doi: 10.1099/mgen.0.000151 [Epub ahead of print]
- Santiviago, C.A., Blondel, C.J., Quezada, C.P., Silva, C.A., Tobar, P.M., Porwollik, S., et al. (2010). Spontaneous excision of the *Salmonella enterica* serovar Enteritidis-specific defective prophage-like element  $\phi$ SE14. *J. Bacteriol.* 192, 2246-2254. doi: 10.1128/JB.00270-09
- Schneider, S., Roessli, D., and Excoffier, L. (2000). *Arlequin: A Software for Population Genetics Data Analysis*, Vol. 2. Geneva: Genetic and Biomedical Laboratory, 2496–2497.
- Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 30, 2068-2069. doi: 10.1093/bioinformatics/btu153
- Timme, R.E., Pettengill, J.B., Allard, M.W., Strain, E., Barrangou, R., Wehnes, C., et al. (2013). Phylogenetic diversity of the enteric pathogen *Salmonella enterica* subsp.

- enterica* inferred from genome-wide reference-free SNP characters. *Genome Biol. Evol.* 5, 2109-2123. doi: 10.1093/gbe/evt159
- Treangen, T.J., Ondov, B.D., Koren, S., and Phillippy, A.M. (2014). The Harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. *Genome Biol.* 15:524. doi: 10.1186/s13059-014-0524-x
- Vernikos, G.S., Thomson, N.R., and Parkhill, J. (2007). Genetic flux over time in the *Salmonella* lineage. *Genome Biol.* 8:R100. doi: 10.1186/gb-2007-8-6-r100
- Wattiau, P., Boland, C., and Bertrand, S. (2011). Methodologies for *Salmonella enterica* subsp. *enterica* subtyping: gold standards and alternatives. *Appl. Environ. Microbiol.* 77, 7877-7885. doi: 10.1128/AEM.05527-11
- Wattiau, P., Van Hesse, M., Schlicker, C., Vander Veken, H., and Imberechts, H. (2008). Comparison of classical serotyping and PremiTest assay for routine identification of common *Salmonella enterica* serovars. *J. Clin. Microbiol.* 46, 4037-4040. doi: 10.1128/JCM.01405-08
- Yachison, C.A., Yoshida, C., Robertson, J., Nash, J.H.E., Kruczkiewicz, P., Taboada, E.N., et al. (2017). The Validation and Implications of Using Whole Genome Sequencing as a Replacement for Traditional Serotyping for a National *Salmonella* Reference Laboratory. *Front. Microbiol.* 8:1044. doi: 10.3389/fmicb.2017.01044
- Yoshida, C.E., Kruczkiewicz, P., Laing, C.R., Lingohr, E.J., Gannon, V.P., Nash, J.H., et al. (2016). The *Salmonella In Silico* Typing Resource (SISTR): An Open Web-Accessible Tool for Rapidly Typing and Subtyping Draft *Salmonella* Genome Assemblies. *PLoS One* 11:e0147101. doi: 10.1371/journal.pone.0147101
- Zhang, S., Yin, Y., Jones, M.B., Zhang, Z., Deatherage Kaiser, B.L., Dinsmore, B.A., et al. (2015). *Salmonella* serotype determination utilizing high-throughput genome sequencing data. *J. Clin. Microbiol.* 53, 1685-1692. doi: 10.1128/JCM.00323-15
- Zou, Q.H., Li, R.Q., Liu, G.R., and Liu, S.L. (2016). Genotyping of *Salmonella* with lineage-specific genes: correlation with serotyping. *Int. J. Infect. Dis.* 49, 134-140. doi: 10.1016/j.ijid.2016.05.029
- Zou, Q.H., Li, R.Q., Wang, Y.J., and Liu, S.L. (2013). Identification of genes to differentiate closely related *Salmonella* lineages. *PLoS One* 8:e55988. doi: 10.1371/journal.pone.0055988

# **Chapter 3. Highly sensitive and specific detection and serotyping of five prevalent *Salmonella* serovars by Multiple Cross Displacement Amplification**

## **3.1 Link to thesis**

*Salmonella* is one of the most common causes of foodborne disease worldwide, including in Australia. The five co-circulating *Salmonella* serovars: Typhimurium, Enteritidis, Virchow, Saintpaul, and Infantis caused over 85% of human *Salmonella* infections in Australia. A simple, rapid, sensitive and specific method to detect *Salmonella* and identify different serovars is essential for public health investigation. Based on the work of Chapter 2, the serovar-specific gene markers obtained from extensive *in silico* genome analysis whose presence or absence can be used to predict a serovar from genomic data. The presence or absence of serovar-specific gene markers from a strain may also be useful in the development of more cost-effective laboratory molecular diagnostics assays to detect them. This prompted me to use a cutting edge molecular assay platform to detect these five serovar-specific gene markers. This work shows clear and concise evidence that a unified approach using serovar-specific gene markers and a common detection assay platform can offer a rapid, accurate and sensitive method for serotyping of common *Salmonella* serovars. This chapter addresses the second aim of this thesis.

I have published this work:

Zhang X, Payne M, Wang Q, Sintchenko V, Lan R. Highly Sensitive and Specific Detection and Serotyping of Five Prevalent *Salmonella* Serovars by Multiple Cross-Displacement Amplification. The Journal of molecular diagnostics : JMD. 2020;22(5):708-19.

I have presented this work at national conference:

Zhang X, Payne M, Wang Q, Sintchenko V, Lan R. Highly Sensitive and Specific Detection and Serotyping of Five Prevalent *Salmonella* Serovars by Multiple Cross-Displacement Amplification. Poster presentation, Australian Society for Microbiology Annual Scientific Meeting 2019.

**Supplemental material for this article can be found at:**

<https://doi.org/10.1016/j.jmoldx.2020.02.006>.

<https://drive.google.com/drive/folders/1DpbvFwt32VMbM38vcmnWGPocRo4hiZYQ?usp=sharing>

**Supplemental material for this article is also listed at Appendix II.**

### 3.2 Abstract

*Salmonella* is a common cause of foodborne disease worldwide including Australia. Over 85% of outbreaks of human salmonellosis in Australia were caused by five *Salmonella* serovars. Rapid, accurate and sensitive identification of *Salmonella* serovars is vital for diagnosis and public health surveillance. Recently, an isothermal amplification technique termed Multiple Cross Displacement Amplification (MCDA) has been employed to detect *Salmonella* at the species level. In the current study, we developed and evaluated seven MCDA assays for rapid detection and differentiation of the five most common *Salmonella* serovars in Australia: Typhimurium, Enteritidis, Virchow, Saintpaul, and Infantis. MCDA primer sets were designed by targeting seven serovar/lineage-specific gene markers identified through genomic comparisons. The sensitivity and specificity of the seven MCDA assays were evaluated using 79 target strains and 32 non-target strains. The assays were all highly sensitive and specific to target serovars with the sensitivity ranged from 92.9% to 100% and the specificity ranged from 93.3% to 100%. The limit of detection of the seven MCDA assays was 50 fg per reaction (10 copies) from pure DNA and positive results were detected in as little as 8 minutes. These seven MCDA assays offer a rapid, accurate and sensitive serotyping method. With further validation in clinically relevant conditions these assays could be used for culture-independent serotyping of common *Salmonella* serovars directly from clinical samples.

**Running title:** MCDA for typing of *Salmonella* serovars

**Funding:** Supported by the National Health and Medical Research Council of Australia project grant.

### 3.3 Introduction

*Salmonella enterica* is one of the most common causes of food-borne disease worldwide including Australia <sup>1,2</sup>. The number of reported human salmonellosis cases in Australia have increased significantly during recent decades <sup>3</sup>. The case rate is estimated to be 185 per 100,000 population per year <sup>4</sup>, with 16,383 cases being notified in 2017, a 30% increase compared with the mean notifications for the previous 10 years (2007-2016) (National Notifiable Diseases Surveillance System of Australia). The most prevalent serovar of *Salmonella* reported in human infections in Australia has been Typhimurium <sup>1,5-7</sup>, followed by Enteritidis, Virchow, Saintpaul, and Infantis <sup>1</sup>. These five co-circulating *Salmonella* serovars are responsible for over 85% of outbreaks of human salmonellosis in Australia <sup>1</sup>.

Traditional culture-based methods for detection of *Salmonella* pathogens are time-consuming, laborious and expensive <sup>8-11</sup>. In recent decades, many culture-independent methods using genetic determinants have offered appealing alternatives to traditional methods. Among these, PCR based techniques (PCR and Real-time PCR) and isothermal amplification techniques, such as LAMP, have been suggested <sup>12-18</sup>. In a recent study, another isothermal amplification technique, termed Multiple Cross Displacement Amplification (MCDA), was reported <sup>19,20</sup>. MCDA employs ten primers, instead of 6 in LAMP or 2 in PCR, to recognize 10 distinct regions, which enhance its specificity and sensitivity.

Comparative genomics leveraging the recent explosion of publicly available genomic data can be used to identify molecular markers for pathogen detection. In our previous study <sup>21</sup> we utilised comparative genomics to identify a panel of 15 serovar/lineage-specific gene markers for typing the 10 most frequent *Salmonella* serovars in Australia. When prevalence of *Salmonella* serovars endemic in Australian was taken into account, the genes listed in that panel can be used as markers for laboratory based typing with an error rate of less than 2.4%.

In this study, we aimed to develop and evaluate MCDA assays targeting serovar/lineage-specific genes for rapid, accurate identification and serotyping of the five *Salmonella* serovars Typhimurium, Enteritidis, Virchow, Saintpaul, and Infantis, which have been dominant in Australia and internationally.

### **3.4 Materials and Methods**

#### **3.4.1 Bacterial strains and Genomic DNA extraction**

A total of 111 strains were used consisting of: 16 *Salmonella* Reference Collection A (SARA) strains <sup>22</sup>; 33 *Salmonella* Reference Collection B (SARB) strains <sup>23</sup> including 9 target serovar strains and 24 non-target serovar strains; 54 *Salmonella* strains from NSW Enteric Reference Laboratory, NSW Health Pathology representing 5 target serovars; and 8 non-*Salmonella* strains representing 8 different species (Supplementary Table 1). The bacterial strains were cultured on nutrient agar at 37°C for 18-24 hours. Genomic DNA was extracted using phenol-chloroform method as described previously <sup>24</sup>.

#### **3.4.2 Design of MCDA primers and the specificities of MCDA products**

Seven serovar/lineage-specific gene markers from our previous study<sup>21</sup> (STM4494, SEN1384, R561\_RS18155, SESV\_RS06060, SESPA\_RS08460, SeSPB\_A1749 and L287\_11788) specific to the five most common *Salmonella* serovars in Australia and internationally (Typhimurium, Enteritidis-clade B, Enteritidis-clade A/C, Virchow, Saintpaul lineage I (Saintpaul-I), Saintpaul lineage II (Saintpaul-II), and Infantis) were selected as targets for MCDA primers design. The sequences of the seven serovar/lineage-specific gene markers are listed in Supplementary Data 1.

Each MCDA primer set consisted of 2 cross primers (CP1 and CP2), 2 displacement primers (F1 and F2), and 6 amplification primers (D1, C1, R1, D2, C2, and R2) <sup>19</sup>. Seven MCDA primers sets were designed using Primer3 online software based on the principle of MCDA recognizing 10 distinct regions to amplify each serovar/lineage-specific gene marker. The specificities of the primers were analysed by NCBI BLAST. OligoAnalyzer





**Chapter 3. Table 1: Primers used for seven Multiple Cross Displacement Amplification assays**

<b>*Primer's name</b>	<b>Sequence (5' → 3')</b>
Typhimurium-F1	AATCGTCGCTCTTCAATATG
Typhimurium-F2	TGTAGCCAGCGTTGTACC
Typhimurium-CP1	GATACGTTTACCGCTGAAGAACTGG- AACCATGCCCCGGTGAATATC
Typhimurium-CP2	TCAGGGAATGATCATTCGTTAGATGC- TAAACAGCATAATCAGCACCTG
Typhimurium-C1	GATACGTTTACCGCTGAAGAACTGG
Typhimurium-C2	TCAGGGAATGATCATTCGTTAGATGC
Typhimurium-D1	TAGCGTGCGGATCATTTCA
Typhimurium-D2	CTTAGCTCCGGCGAACAT
Typhimurium-R1	CCTGGATGAATTTTCAGCTTC
Typhimurium-R2	GCAACGTGTCCTACTGGAT
Enteritidis-clade B-F1	ATAACACTTACGGAGCTGAG
Enteritidis-clade B-F2	TCGTAACGACGTACCTCAC
Enteritidis-clade B-CP1	CCACAACGTTCTGCCTTGTCCAAGGATGACGGG GTTAACCATT
Enteritidis-clade B-CP2	GCTTATCGTGCCTGGAAGAAACAGCGTCAGGCA GCTTCCAAATC
Enteritidis-clade B-C1	CCACAACGTTCTGCCTTGTCCA
Enteritidis-clade B-C2	GCTTATCGTGCCTGGAAGAAACAG
Enteritidis-clade B-D1	GTAGTGGCGGGTCAATA
Enteritidis-clade B-D2	GAAAGTGGACGCTGACCT
Enteritidis-clade B-R1	TGCCCCGCTGGTACACAT
Enteritidis-clade B-R2	GATTTTCCCGTCAGAAGAG
Enteritidis-clade A/C-F1	TTTCATTATAGGGCAGGGA
Enteritidis-clade A/C-F2	CTGTCACAATCAAATAATGA
Enteritidis-clade A/C-CP1	GTGACACGAAATGAATGAGTCCAATCGTCTTGA GATTATAGTTACTCTTG
Enteritidis-clade A/C-CP2	TGCGAGTAGGTATTTATAAGGTTGAGTCATGTA TATTAAACTCTGGTC
Enteritidis-clade A/C-C1	GTGACACGAAATGAATGAGTCCAATC
Enteritidis-clade A/C-C2	TGCGAGTAGGTATTTATAAGGTTGAG
Enteritidis-clade A/C-D1	CGAAAATCCGAATTCCTCC
Enteritidis-clade A/C-D2	ACTCATCTTATCTGGAATGG
Enteritidis-clade A/C-R1	CAACAGATCACCTTCATCA
Enteritidis-clade A/C-R2	TGTTGGGTGAGCAAAAAGG
Virchow-F1	TCATTATTAGACCAATCTGC
Virchow-F2	TTCGTTTGCTGATTCCATG

Virchow-CP1	GTGCTGAAACTTTTATTTATGCTTGGAATTGA CCAGTCGGTTAAGGC
Virchow-CP2	GCCAGCACAAATGAATACTGTATGGCAACGGG ATCCTATTC
Virchow-C1	GTGCTGAAACTTTTATTTATGCTTGGG
Virchow-C2	GCCAGCACAAATGAATACTGT
Virchow-D1	AACTTTCGCGTTGTGAGCT
Virchow-D2	CTGGATCTTAAATAGTCATC
Virchow-R1	ATTTTAGGTGGCACCCATC
Virchow-R2	TATGTTGTGGCATATGATGG
Saintpaul-I-F1	TCAGACTGAAGACCAGCTT
Saintpaul-I-F2	TAGCATCTTTAGTACCAGC
Saintpaul-I-CP1	TCCACTGAGCGGAAAAATGCCAGAAAGCTAAA AGGATATACGGG
Saintpaul-I-CP2	CTGGATGGCTCTCTGGTGCTTCCGTAGCTTGCA GCGTTTC
Saintpaul-I-C1	TCCACTGAGCGGAAAAATGCCAG
Saintpaul-I-C2	CTGGATGGCTCTCTGGTGCT
Saintpaul-I-D1	GCGACATTGGGTGTAATC
Saintpaul-I-D2	GATAAAATAACGTGGCTGG
Saintpaul-I-R1	GCGAATAGCGAACTCACT
Saintpaul-I-R2	TGAGCGGGATAGTAAGAAG
Saintpaul-II-F1	TTATTACCAGTGCCGCGAT
Saintpaul-II-F2	TGTAGCCAGCGTTGTACC
Saintpaul-II-CP1	CACCACGTTTTTAGGGCTGATGAAGCGGGCTCT TTTAATGCTAAGT
Saintpaul-II-CP2	GACATTTCTCACCTTCCAGGGCTTCCGTATCA AGGTTATGGG
Saintpaul-II-C1	CACCACGTTTTTAGGGCTGATGAAG
Saintpaul-II-C2	GACATTTCTCACCTTCCAGGG
Saintpaul-II-D1	GAAATTTCTCGGAGCCAGT
Saintpaul-II-D2	TGAAGGGATCCTGTTTTCTG
Saintpaul-II-R1	CATAACAATGCTTTTGTTGCC
Saintpaul-II-R2	TACCTGATCGATGACACTC
Infantis-F1	TTATGGCTGACAACGAGAG
Infantis-F2	ATCCAGGTCAAACGCTTGC
Infantis-CP1	TCCGACTCTGCGTTAACGATGCTATTCATCCTG ATGTCGCTC
Infantis-CP2	TCAAGGCATCGAAAACCTGATCCTGACTGTAGA AAGCACAAACACC
Infantis-C1	TCCGACTCTGCGTTAACGATG
Infantis-C2	TCAAGGCATCGAAAACCTGATCCT

Infantis-D1	ACGACCTCATTCTGCC
Infantis-D2	TACCGGTGTGACTACCAG
Infantis-R1	GTTCGGTAAACGAGAAAGC
Infantis-R2	TGAGATGATCCTTCGTGC

\*: Typhimurium: STM4494; Enteritidis-clade B: SEN1384; Enteritidis-clade A/C: R561\_RS18155; Virchow: SESV\_RS06060; Saintpaul-I: SESPA\_RS08460; Saintpaul-II: SeSPB\_A1749; Infantis: L287\_11788.

software was used for primer dimer and secondary structure investigation. The sequences and locations of seven MCDA primers sets are presented in Table 1 and Figure 1.

The specificity of seven MCDA products (i.e. presence only in the targeted serovars) was also examined *in silico* by searching the products against a diverse set of 2258 *Salmonella* species genomes from the identification data group in our previous study <sup>21</sup> using BLASTN with default settings.

#### **3.4.3 The initial evaluation of the seven MCDA assays**

To evaluate the seven MCDA primers sets, *Salmonella* species specific gene *invA* MCDA assay <sup>20</sup> was utilized as positive control. The MCDA reactions were performed on a Corbett Rotor-Gene 6000 Real Time PCR Machine with the WarmStart LAMP DNA amplification Kit (New England BioLabs, Sydney, Australia) in a total volume of 10 µL reaction mixture incubated at 63°C for 60 min and then heated at 95°C for 5 min to stop the amplification. Real-time LAMP Fluorescent dye (FD) measurement was used to monitor the MCDA amplification every minute.

The final 10 µL MCDA reaction mixtures contained the primers' concentration as previously described <sup>20</sup>: 2.4 µM each of cross primers CP1 and CP2, 0.4 µM each of displacement primers F1 and F2, 1.2 µM each of amplification primers R1, R2, D1 and D2, 0.8 µM each of amplification primers C1 and C2. The reaction mixture was consisted of 5 µL 2X WarmStart LAMP Master Mix, 0.2 µL 5X FD, 1.2 µl MCDA primers mixture, 2.6 µL Milli-Q water and 1 µL DNA template.

#### **3.4.4 Evaluation of the limit of detection of the MCDA assays in pure culture**

Limit of detection (LoD) was defined as the lowest genomic DNA level where all 6 replicates were detected by each MCDA assay on the condition that its detection time was faster than *invA* MCDA assay. To demonstrate the efficiency of seven MCDA assays, each novel MCDA assay together with the existing *Salmonella invA* MCDA assay was analysed for LoD. The genomic DNA template from seven target strains belonging to their respective serovars (SARA14, Typhimurium; L2376, Enteritidis-clade B; L2380, Enteritidis-clade A/C; L2349, Virchow; SARA28, Saintpaul-I; SARB56, Saintpaul-II; L2385, Infantis) were serially diluted (5 ng, 500 pg, 50 pg, 5 pg, 500 fg, 50 fg, 25 fg, 12.5

fg, 6.25 fg per microliter) to determine the LoD. One replicate for the dilutions 5 ng, 500 pg, 50 pg and 5 pg and three replicates for the dilutions 500 fg, 50 fg, 25 fg, 12.5 fg, 6.25 fg were tested with two independent runs.

The detection time was defined as the time at which the fluorescence signal doubled the value of the baseline. GraphPad Prism was used to perform statistical analyses to show the relationship of detection times and dilutions between each MCDA assay and *invA* MCDA assay.

#### **3.4.5 Evaluation of the sensitivity and specificity of seven MCDA assays in pure culture**

The sensitivity of each MCDA assay was defined as the percentage of strains of a targeted serovar being detected as positive. The sensitivity was evaluated with genomic DNA templates from 79 target strains (Typhimurium n=11, Enteritidis n=24, Virchow n=10, Saintpaul n=20, Infantis n=14) under the same conditions described above. Specificity was defined as the percentage of strains from non-targeted serovars being detected as negative and ideally should be 100%. A panel of 30 strains including 24 non-target strains from SARB collection representing 24 serovars and target strains from the other 6 assays were used to analyse the specificity of MCDA assays within *Salmonella* species. An additional 8 non-*Salmonella* strains were tested for specificity of seven MCDA assays with other species.

All strains were tested in duplicate at 500 pg / $\mu$ L level with two independent runs. Both replicates with a kinetic graph within 30 minutes incubation time were considered as positive amplification.

#### **3.4.6 Phylogenetic analyses**

Whole-genome sequencing (WGS) of 4 SARB Enteritidis strains was performed by Illumina NextSeq (Illumina, Scoresby, VIC, Australia). DNA libraries were constructed using Nextera XT Sample preparation kit (Illumina Inc., San Diego, CA, USA) and sequenced using the NextSeq sequencer (Illumina Inc.). FASTQ sequences were deposited in the National Center for Biotechnology Information (NCBI) Short Read Archive under the BioProject PRJNA552918. Raw reads for these strains were de novo

assembled using SPADIS v3.10.1 assembler with default settings<sup>25</sup>. The serovar of the 4 assembled SARB Enteritidis genomes was predicted by *Salmonella In Silico* Typing Resource (SISTR)<sup>26</sup> and SeqSero<sup>27</sup>. The sequences of flagellin gene *fliC* were extracted from SeqSero<sup>27</sup> database. The genome of Infantis SARB27 strain was downloaded from the NCBI GenBank (RefSeq assembly accession: GCF\_000230875.1).

To investigate the phylogeny of observed false negative strains and false positive strains with related serovars, parsnp v1.2<sup>28</sup> with default parameters was used to generate phylogenetic trees from genomes. The tree was visualised using Figtree v1.4.3<sup>29</sup>. A phylogenetic tree of the flagellin encoding *fliC* gene sequence was constructed using Mega X with default parameters by Maximum Parsimony method with 500 bootstrap replicates<sup>30</sup>.

## 3.5 Results

### 3.5.1 Selection of serovar specific genes for MCDA products targeting five serovars

One marker each for monophyletic serovars was selected. STM4494, SESV\_RS06060 and L287\_11788 were selected for Typhimurium, Virchow and Infantis, respectively, which were identified and confirmed as serovar-specific markers previously<sup>21</sup>. For polyphyletic serovar Saintpaul (Saintpaul-lineage I and Saintpaul-lineage II) and paraphyletic serovar Enteritidis (Enteritidis-clade B and Enteritidis-clade A/C) more than one marker was required to identify the different lineages of the serovars. One marker each was selected for Saintpaul-lineage I and Saintpaul-lineage II (SESPA\_RS08460 and SeSPB\_A1749 respectively). SEN1384 was selected for Enteritidis-clade B and R561\_RS18155 was selected for Enteritidis-clade A/C. A total of seven MCDA assays based on these markers were designed to detect these five serovars.

We examined the seven MCDA products *in silico* using BLASTN against 2258 genomes used in our previous study<sup>21</sup> to confirm the specificities of each MCDA products. The BLASTN results showed that each MCDA product was found in the same genomes as the respective genes with no additional false positives in agreement with our previous study (Supplementary Table 2). In all cases false positive genomes were from serovars

that were rare in human infections and are not expected to be major limitations on the applicability of these assays in a clinical setting.

### **3.5.2 Evaluation of limit of detection (LOD) of the seven MCDA assays in pure culture**

The LOD of each of the seven MCDA assays designed in this study and the species identification *invA* MCDA assay<sup>20</sup> was performed on serially diluted genomic DNA. The amplification curves of the eight MCDA assays are shown in Figure 2. The seven serovar specific target strains can be detected within 8 minutes incubation time at highest concentration tested (5 ng/μL) in all seven MCDA assays. While the detection time of *invA* MCDA assay at this same concentration was 12 minutes.

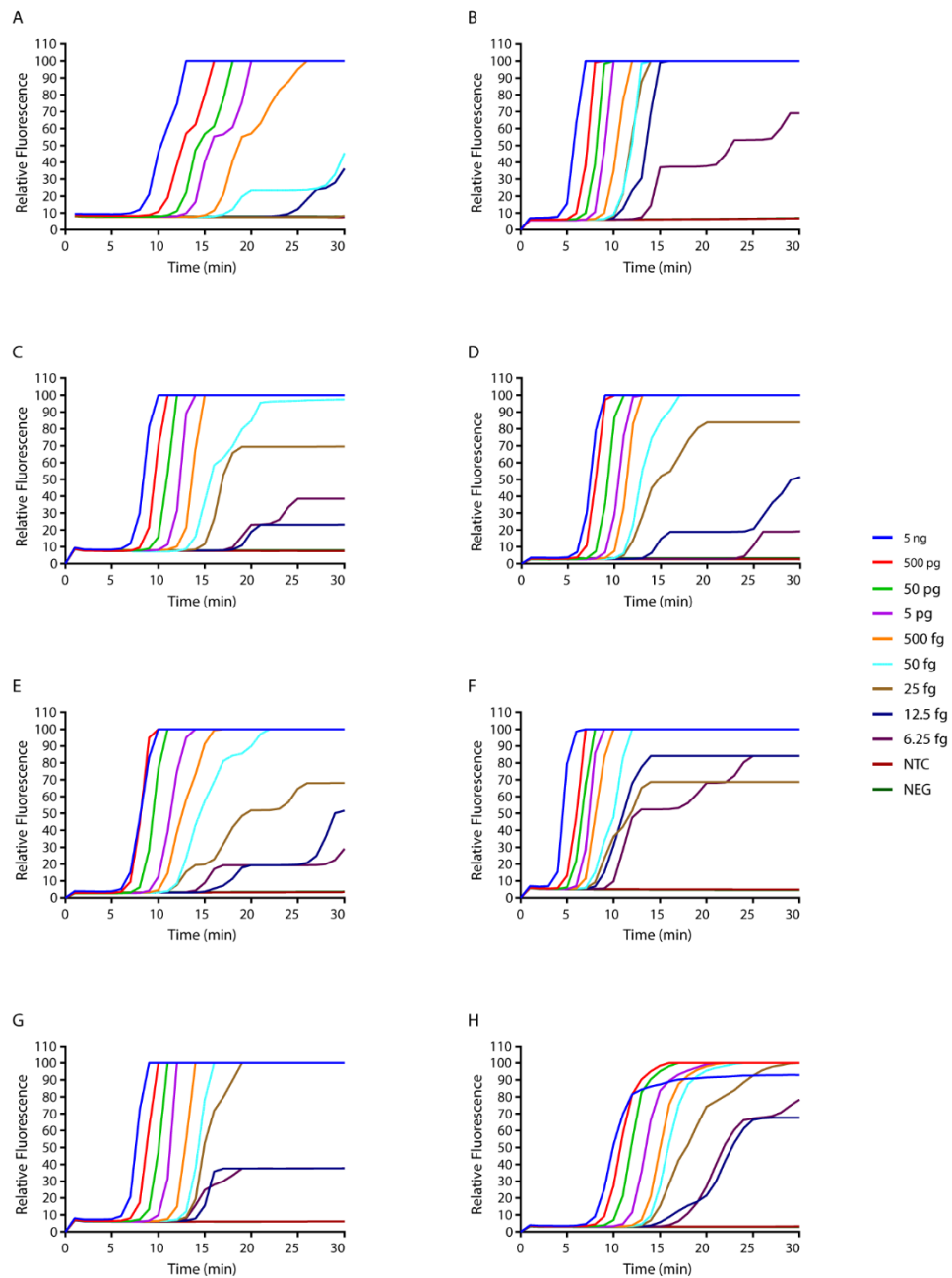
Enteritidis-clade B, Enteritidis-clade A/C, Virchow and Infantis MCDA assays had LoD of 50 fg/μl (10 copies). The LoD for Saintpaul MCDA assays (Saintpaul-I and Saintpaul-II) was 25 fg/μl (5 copies) and Typhimurium MCAD assay 12.5 fg/μl (2.5 copies), respectively. Successful detection of targets was also observed at even lower concentrations for seven MCDA assays not for all 6 replicates. For example, Enteritidis-clade A/C with R561\_RS18155 MCDA assay detected positive amplifications at 25 fg/μL and 12.5 fg/μL DNA level with 5 out of 6 replicates and 3 out of 6 replicates, respectively.

In order to provide a way to distinguish low copy number positive results from false positives we compared the LoD curves of the *invA* MCDA assay and the seven MCDA assays designed here. In all assays and concentrations the *invA* MCDA assay positive results were observed after the serovar specific result (Figure 3). Therefore, *invA* assay acted as a benchmark for serovar specific MCDA assay. If the *invA* assay provides a positive result after the serovar specific result the serovar specific results can be trusted. If *invA* assay is positive before the serovar specific assay or negative completely, the result is likely to be a false positive.

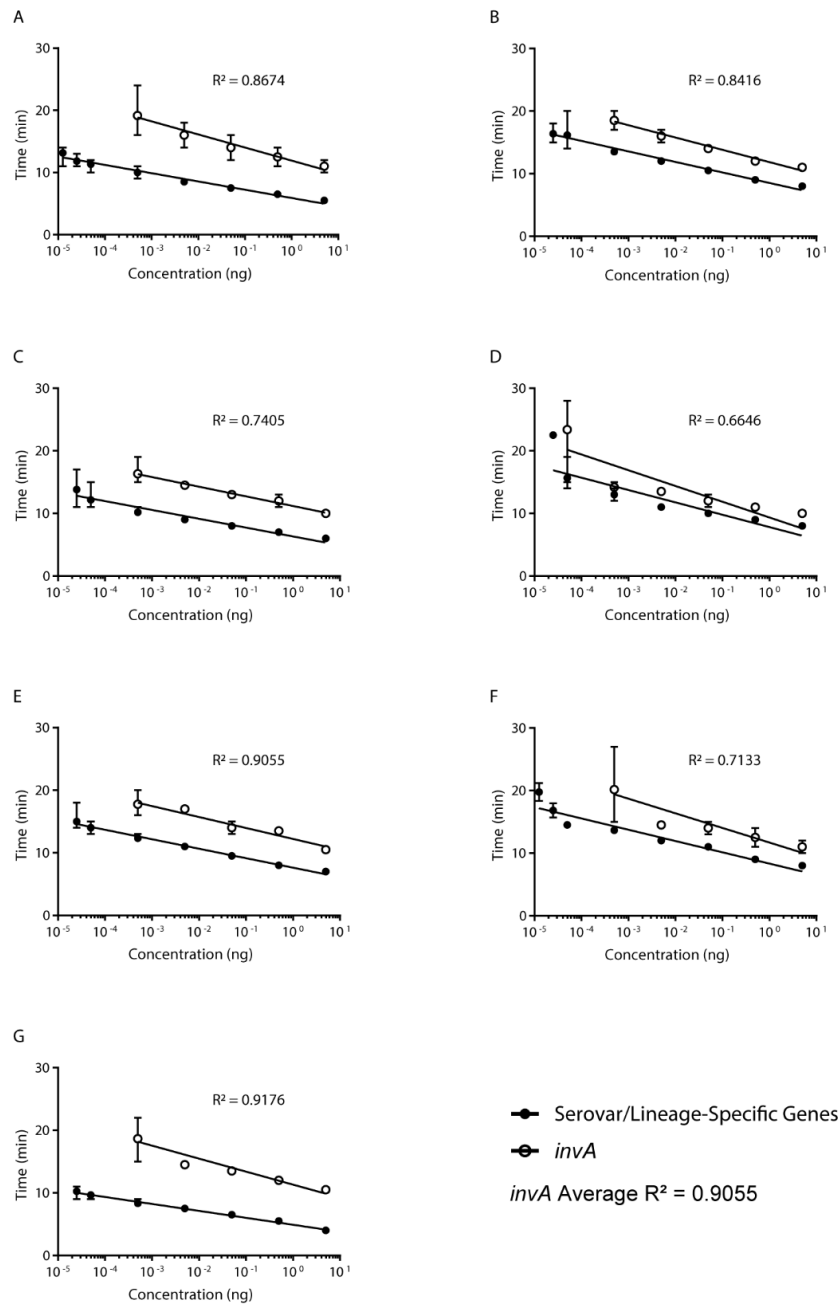
### **3.5.3 Evaluation of the sensitivity of seven MCDA assays in pure culture**

The sensitivity of each MCDA assay was determined against target strains listed in Supplementary Table 1. All target strains were successfully amplified by respective MCDA assays, except for Infantis strain SARB27 and four Enteritidis strains (SARB16,





**Chapter 3. Figure 2: Limit of Detection (LoD) amplification curves of seven Multiple Cross Displacement Amplification (MCDA) assays.** LoD amplification curves of seven MCDA assays generated by GraphPad Prism were listed. A, *invA*; B, Typhimurium; C: Enteritidis-clade B; D: Enteritidis-clade A/C; E: Virchow; F: Saintpaul-I; G: Saintpaul-II; H: Infantis. Curves for each concentration of DNA were marked in the figure. The concentration started with 5 ng per reaction and decreased from left to right as per the color key. Each point on each line was the average of relative fluorescence of at least 6 replicates. LoD was defined as the lowest concentration at which all replicates produced final relative fluorescence of over 90%. LoD for each of the assays was as follow. 50 fg (10 copies): Enteritidis-clade B, Enteritidis-clade A/C, Virchow and Infantis. 25 fg (5 copies): Saintpaul-I and Saintpaul-II. 12.5 fg (2.5 copies): Typhimurium. NTC: No Template Control; NEG: *E. coli* K12.



**Chapter 3. Figure 3: Standard curves of seven Multiple Cross Displacement Amplification (MCDA) assays based on average detection times and serial dilutions.** Standard curves of average detection times based on serial dilutions for each of the seven MCDA assays. A: Typhimurium; B: Enteritidis-clade B; C: Enteritidis-clade A/C; D: Virchow; E: Saintpaul-I; F: Saintpaul-II; G: Infantis. Solid circle indicates seven MCDA assays and hollow circle indicates *invA* MCDA assay. Error bars represent the range of all replicates at each concentration. The average detection times for each MCDA assay were linear along the dilution gradient and had a significant correlation ( $P < 0.001$ ) with linear standard curves. Parallel lines between each MCDA assay and *invA* assay were observed. In all assays and concentrations the *invA* positive results were observed after the serovar specific results.

SARB17, SARB18, and SARB19). Among these 5 target strains, Infantis strain SARB27 was negative in the Infantis MCDA assay. For the 4 Enteritidis strains, only SARB18 was detected by Enteritidis-clade A/C MCDA assay.

We performed BLASTN searches of the Infantis specific gene marker against SARB27 genome sequence downloaded from NCBI and BLASTN results indicated that SARB27 lacked Infantis specific gene marker. We also conducted BLASTN searches of the two Enteritidis specific gene markers against the 4 SARB Enteritidis genomes we sequenced we found that 3 SARB Enteritidis strains (SARB16, SARB17, and SARB19) lacked both Enteritidis gene markers while SARB18 contained Enteritidis-clade A/C gene marker R561\_RS18155.

In order to confirm the serovar's identity, we performed the serovar prediction of the 4 assembled SARB Enteritidis genomes using SISTR<sup>26</sup> and SeqSero<sup>27</sup>. SARB16 and SARB19 were assigned to serovars Duisburg and Emek, respectively. SARB17 and SARB18 were predicted as serovar Enteritidis. Phylogenetic tree of genomes constructed using parsnp revealed that SARB18 was clustered with Enteritidis-clade A/C, whereas SARB16, SARB17 and SARB19 were grouped with other serovars (Supplementary Figure 1). A SNP-based phylogenetic tree of *fliC* (Supplementary Figure 2) indicated that Enteritidis-clade A/C and Enteritidis-clade B were identical. In contrast, SARB17 *fliC* was not grouped with Enteritidis *fliC* and they differed by 3 SNPs with one being a non-synonymous SNP. We concluded that SARB17 does not represent *S. Enteritidis*.

Combined with the serovar prediction results and phylogenetic analysis, SARB18 was assigned to Enteritidis-clade A/C while SARB16, SARB17 and SARB19 were assigned to other serovars. Therefore SARB16, SARB17 and SARB19 were excluded from the target strains of Enteritidis-clade A/C and Enteritidis-clade B. The sensitivity of each MCDA assay is defined as positive rate of targeted serovar strains. The sensitivity of seven MCDA assays varied from 92.9% to 100% (Table 2).

#### 3.5.4 Evaluation of the specificity of seven MCDA assays in pure culture

The specificity of each MCDA assay was evaluated against 8 non-*Salmonella* strains and 30 *Salmonella* strains (Supplementary Table 1). The specificity of seven MCDA assays with other species were 100% while the specificity of seven MCDA assays with non-targeted *Salmonella* strains were varying from 93.3% to 100% (Table 2). Typhimurium and Enteritidis-clade A/C MCDA assays produced 2 false positive results each. One Derby strain (SARB9) and one Infantis strain (SARB27) were detected by Typhimurium MCDA assay, which led to the specificity of 93.3% for Typhimurium MCDA assay. One strain each for Dublin and Gallinarium (SARB13 and SARB21 respectively) were amplified by Enteritidis-clade A/C MCDA assay, therefore, the specificity of Enteritidis-clade A/C MCDA assay was 93.3%.

BLASTN result showed that Infantis SARB27 contained Typhimurium gene marker STM4494. Phylogenetic tree constructed using parsnp indicated that SARB27 did not cluster with Typhimurium and was not part of the main Infantis clade (Supplementary Figure 3).

### 3.6 Discussion

*Salmonella* is a common cause of foodborne illness in Australia <sup>1</sup>. Rapid, accurate and sensitive detection and identification of *Salmonella* serovars have been essential for clinical diagnosis and public health surveillance. In recent decades, PCR-based and real-time PCR techniques have frequently been used to detect and differentiate *Salmonella* serovars <sup>10, 16, 31-38</sup>. LAMP was also used as an alternative method for rapid and sensitive detection of specific gene targets for identification of *Salmonella* <sup>13, 39-41</sup>. A few LAMP assays can also differentiate Typhimurium and Enteritidis from other *Salmonella* serovars <sup>12, 14, 15, 40, 42-46</sup>. The limit of detection for the reported *Salmonella* LAMP assays ranged from 5 fg to 5.6 ng genomic DNA per reaction in pure-culture and the comparison between LAMP and PCR or real-time PCR performed in the same study showed that LAMP was 10 to 10,000-fold more sensitive <sup>47</sup>. Another isothermal amplification technique, MCDA, was developed in 2015 with at least 160-fold and 16-fold higher analytical sensitivity than PCR and LAMP, respectively <sup>19</sup>. MCDA has been employed to detect *Salmonella* at the species level but not to the serovar level. MCDA can detect

*Salmonella* at 6.25 fg genomic DNA per reaction, at least 400-fold more sensitive than real-time PCR <sup>20</sup>.

With the increasing uptake of culture-independent direct testing for the diagnosis of *salmonella* enterocolitis, culture-independent serovar detection and identification targeting serovar-specific gene markers would be useful as current serotyping is mostly performed on DNA from purified isolates. A number of genes STM4493, STM4495, STM4497, *typh*, *lygD* (SEN1383), *sdfI*, *safA*, *prot6E* and *sefA*, have been used to develop PCR-based methods and LAMP assays for typing Typhimurium and Enteritidis <sup>12, 14, 36, 42, 46, 48</sup>. In our previous study<sup>21</sup>, genes STM4493, STM4494 and STM4497 were identified as potential of Typhimurium gene markers but were present in Infantis SARB27. However, with genomic data of 2258 genomes, the specificities of gene markers STM4493 and STM4497 were 93.22% and 92.56% respectively, while the specificity of gene marker STM4494 was 94.11%, slightly higher than STM4493 and STM4497. A novel PCR was also developed for identification of Infantis based on flagellin *fljB* gene <sup>49</sup> and Virchow-specific primers was designed for the detection of Virchow <sup>50</sup>.

Our results indicate that a set of MCDA assays targeting seven serovar/lineage-specific gene markers can rapidly detect and serotype the five most common and clinically relevant *Salmonella* serovars. Seven serovar/lineage-specific gene markers, STM4494, SEN1384/R561\_RS18155, SESV\_RS06060, SeSPB\_A1749/SeSPA\_A1352 and L287\_11788 <sup>21</sup> were used to develop Typhimurium, Enteritidis-clade B/Enteritidis-clade A/C, Virchow, Saintpaul-I/Saintpaul-II, and Infantis MCDA assays respectively. The initial evaluation of the seven MCDA assays were accomplished by utilizing *Salmonella* species specific gene *invA* MCDA assay <sup>20</sup> as positive control. Multiple means can be used to display the correct amplification of MCDA <sup>19</sup>. In our study, real-time fluorescence measurement was used to detect the seven MCDA products. The seven MCDA assays were very sensitive with a LoD of 50 fg (10 copies) with pure DNA and the assays were rapid with a result detectable within 8 minutes at the highest concentration tested. The assays also were highly specific to target serovars and the positives reactions were monitored in a real-time format. Our assays provided a unified approach using serovar-specific gene markers obtained from extensive *in silico* genome analysis <sup>21</sup> and a

common MCDA assay platform <sup>19</sup>. These gene markers can also be used to develop assays on other platforms.

The speed and LoD of the MCDA assay compared favourably with published LAMP assays. Previous studies <sup>12, 14</sup> evaluated LAMP assays for detection of Typhimurium and Enteritidis by targeting STM4497 and *safA* genes respectively. The fastest detection time in these studies was around 24 min and the LoD were 4.38 pg/μL for Typhimurium and 1.44 pg/μL for Enteritidis in pure culture. By comparison our Typhimurium MCDA targeting STM4494 and Enteritidis-clade B MCDA targeting SEN1384 were nearly 15 minutes faster than LAMP. Additionally, these results were produced with 50 fg/μL of pure DNA, at least 87-fold more sensitive than LAMP for Typhimurium MCDA assay and 29-fold more sensitive than LAMP for Enteritidis-clade B MCDA assay.

MCDA assays could also produce positive results at even lower concentrations of 6.25 fg/μL (1.25 copies) genomic DNA within 28 minutes although these results were not as consistent. This inconsistency may be due to the extremely low copy number of the sample resulting in no template being present by chance during template sampling. Inconsistent amplification with very few copies of the template within 28 minutes indicated that any amplification later than 28 minutes incubation time was unreliable. Therefore a 30 minutes incubation time was used to set the cut-off value for evaluation of sensitivity and specificity of the seven MCDA assays.

In some reactions, amplification can occur at timepoints that most likely contained very few copies of the template. These amplifications may be due to a real (true positive) detection of very dilute target DNA or may be caused by inefficient amplification of non-target DNA (false positive). To differentiate between these options the *invA* MCDA assay can be used as an outer limit on the amplification time of a true positive result. The sample will only be considered successfully serotyped by our newly designed seven MCDA assays when it is positive to both *invA* and the serovar specific target and the amplification of target occurs before *invA*.

**Chapter 3. Table 2: The sensitivity and specificity (%) of the seven MCDA assays**

Target Serovars	Specific Gene <sup>21</sup>	Target strains	Non-target strains n=30	Non- <i>Salmonella</i> strains n=8	Sensitivity*	Specificity* (Within <i>Salmonella</i> )	Specificity* (Non- <i>Salmonella</i> )
Typhimurium	STM4494	11/11	2/30‡	0	100	93.3	100
Enteritidis-clade B	SEN1384	9/9	0	0	100	100	100
Enteritidis-clade A/C	R561_RS18155	12/12	2/30¶	0	100	93.3	100
Virchow	SESV_RS6060	10/10	0	0	100	100	100
Saintpaul-I	SeSPB_A1352	18/18	0	0	100	100	100
Saintpaul-II	SeSPA_A1749	2/2	0	0	100	100	100
Infantis	L287_11788	13/14†	0	0	92.9	100	100

\*: Sensitivity: (No of positive of target strains) / (No of total target strains). Specificity: 1 – (No of positive of non-target strains) / (No of total non-target strains). †: Infantis strain SARB27 was negative to Infantis MCDA assay. ‡: False positive strains: SARB9 (Derby) and SRAB27 (Infantis). ¶: False positive strains: SRAB13 (Dublin) and SARB21 (Gallinarum).

MCDA: Multiple Cross Displacement Amplification.

Three strains SARB16, SARB17, and SARB19 from the *Salmonella* reference collection B which was assembled during the early 1980s<sup>23</sup> were serotyped as Enteritidis. However, SARB16 and SARB19 were assigned as others serovars by WGS based serovar prediction methods SISTR and SeqSero. These three strains were not closely related to Enteritidis on a genome based phylogenetic tree. Previous studies also showed that SARB17 and SARB19 were distantly related to the vast majority of Enteritidis isolates on phylogenetic trees<sup>51, 52</sup>. SNP-based phylogenetic trees of flagellin gene *fliC* indicated that the SARB17 *fliC* gene was distinct from Enteritidis. Consequently SARB16, SARB17 and SARB19 were excluded as target strains for Enteritidis. The genomic signatures that were targeted by the MCDA assays can provide more useful serovar identification than traditional serotyping, especially for those strains with the same serovar but very different evolutionary history.

All 7 MCDA assays had high overall sensitivity ranged from 92.9% to 100%. A false negative result only occurred in SARB27 in the Infantis MCDA assay. Infantis SARB27 which was isolated in Senegal was distantly related to the globally distributed Infantis<sup>23</sup>. All pure culture from clinical strains were correctly detected and identified by the Infantis MCDA assay.

It should be noted that we only used two strains for evaluation the sensitivity of Saintpaul-II MCDA assay. We were unable to test more strains as Saintpaul lineage II is a rare minor lineage, making strain acquisition difficult. From our previous genomic analysis<sup>21</sup> and a further analysis of 291 genomes, the lineage has low diversity and the Saintpaul-II specific gene has 100% *in silico* specificity and sensitivity. Therefore, the assay most likely will be effective in detecting Saintpaul-II isolates.

The false positive Derby strain in the Typhimurium assay is expected based on previous genomic analysis in our previous study which showed that Derby was a potential false positive of the Typhimurium gene marker STM4494<sup>21</sup>. We estimated an ‘Australian potential false positive rate’ for Derby detection in the Typhimurium assay by combining the Derby frequency in Australian human infections and the rate of false positives from Derby genomes. This potential false positive rate was less than 0.41% in human infections in Australia<sup>21</sup>. Therefore, Derby should not be a major limitation to the application of



this assay to human clinical samples in culture independent *Salmonella* serotyping. For veterinary and agricultural samples, the false positive rate may increase, depending on the prevalence of Derby in the source animal population. The prevalence (11.2%) of Derby in pigs and pork in New South Wales<sup>53</sup> would result in a potential false positive rate of 3.1%. However, the routine surveillance of *Salmonella* infections in animal population is limited in Australia<sup>53</sup>, which makes accurate determination of likely false positive rates difficult.

The false positives in the Enteritidis-clade A/C MCDA assay from serovars Dublin and Gallinarum were also expected from previous genomic analysis<sup>21</sup>. Dublin is one of the most prevalent serovars in cattle<sup>54</sup>. However the frequency of Dublin in human infections is less than 1.5% in Australia. The rate of genome based false positives of Dublin was 2.78% for Enteritidis-clade A/C gene marker R561\_RS18155<sup>21</sup>. The Australian potential false positive rate for Dublin in the Enteritidis-clade A/C MCDA assay would be less than 0.04% in human infections. Amongst cattle and beef products (New South Wales prevalence of 33.4%), the potential false positive rate would be 0.93%. Gallinarum is restricted to poultry reservoir in many developing countries<sup>37, 55-58</sup> and remains rare in Australia. The potential false positive rate with Gallinarum in Enteritidis-clade A/C MCDA assay of human samples is therefore negligible.

In conclusion, we developed seven MCDA assays to amplify target gene markers successfully from the five most frequent *Salmonella* serovars in Australia. These assays demonstrated a LoD of 50 fg per reaction (10 copies of target DNA) from pure culture and were specific to the target serovars. The assay is time efficient, isothermal and can provide test results in as little as 8 minutes. The MCDA assays developed offer a rapid, accurate and sensitive serotyping method, which will be useful also for culture-independent serotyping of common *Salmonella* serovars directly from clinical samples. The performance of the MCDA assays warrants further validation on clinical and environmental samples.

### **3.7 Acknowledgements**

The authors thank Peter Howard from the NSW Enteric Reference Laboratory, NSW Health Pathology for providing access to examples of clinical strains. Funding from the

National Health and Medical Research Council to RL and VS is also gratefully acknowledged.

### 3.8 Supplemental Data

Supplemental material for this article can be found at  
<https://doi.org/10.1016/j.jmoldx.2020.02.006>.

### 3.9 References

1. Ford L, Moffatt CRM, Fearnley E, Miller M, Gregory J, Sloan-Gardner TS, Polkinghorne BG, Bell R, Franklin N, Williamson DA, Glass K, Kirk MD: The Epidemiology of Salmonella enterica Outbreaks in Australia, 2001–2016. *Front Sustain Food Syst* 2018, 2:86
2. Majowicz SE, Musto J, Scallan E, Angulo FJ, Kirk M, O'brien SJ, Jones TF, Fazil A, Hoekstra RM, diseases Ci: The global burden of nontyphoidal Salmonella gastroenteritis. *Clin Infect Dis* 2010, 50: 882-889
3. Ford L, Glass K, Veitch M, Wardell R, Polkinghorne B, Dobbins T, Lal A, Kirk MDJPO: Increasing incidence of Salmonella in Australia, 2000-2013. *PLoS One* 2016, 11:-0163989
4. Kirk M, Ford L, Glass K, Hall GJEID: Foodborne illness, Australia, circa 2000 and circa 2010. *Emerg Infect Dis* 2014, 20:1857-1864
5. OzFoodNet-Working-Group: Monitoring the incidence and causes of diseases potentially transmitted by food in Australia: annual report of the OzFoodNet Network, 2009. *Commun Dis Intell Q Rep* 2010, 34: 396-426
6. OzFoodNet-Working-Group: Monitoring the incidence and causes of diseases potentially transmitted by food in Australia: annual report of the OzFoodNet network, 2010. *Commun Dis Intell Q Rep* 2012, 36: E213-E241
7. OzFoodNet-Working-Group: Monitoring the incidence and causes of diseases potentially transmitted by food in Australia: Annual report of the OzFoodNet network, 2011. *Commun Dis Intell Q Rep* 2015, 39: E236-E264
8. ISO-1: Microbiology of the food chain-Horizontal method for the detection, enumeration and serotyping of Salmonella-Part 1: Detection of Salmonella spp. Geneva, Switzerland: International Organiza-tion for Standardization, 2017

9. Andrews WH, Jacobson A, Hammack T: Bacteriological Analytical Manual (BAM). Chapter 5 Salmonella. Washington, DC: U.S. Food and Drug Administration, 2011
10. Hein I, Flekna G, Krassnig M, Wagner MJJomm: Real-time PCR for the detection of Salmonella spp. in food: an alternative approach to a conventional PCR system suggested by the FOOD-PCR project. J Microbiol Methods 2006, 66:538-547
11. Josefsen MH, Krause M, Hansen F, Hoorfar JJAEM: Optimization of a 12-hour TaqMan PCR-based method for detection of Salmonella bacteria in meat. Appl Environ Microbiol 2007, 73:3040-3048
12. Azinheiro S, Carvalho J, Prado M, Garrido-Maestu AJFiSFS: Evaluation of Different Genetic Targets for Salmonella enterica Serovar Enteritidis and Typhimurium, Using Loop-Mediated Isothermal AMPLification for Detection in Food Samples. Front Sustain Food Syst 2018, 2:5
13. Domesle KJ, Yang Q, Hammack TS, Ge BJlJofm: Validation of a Salmonella loop-mediated isothermal amplification assay in animal food. Int J Food Microbiol 2018, 264:63-76
14. Garrido-Maestu A, Fuciños P, Azinheiro S, Carvalho J, Prado MJFC: Systematic loop-mediated isothermal amplification assays for rapid detection and characterization of Salmonella spp., Enteritidis and Typhimurium in food samples. Food Contr 2017, 80:297-306
15. Hu L, Ma LM, Zheng S, He X, Hammack TS, Brown EW, Zhang GJFc: Development of a novel loop-mediated isothermal amplification (LAMP) assay for the detection of Salmonella ser. Enteritidis from egg products. Food Contr 2018, 88:190-197
16. Kasturi KN, Drgon TJAEM: Real-time PCR method for detection of Salmonella spp. in environmental samples. Appl Environ Microbiol 2017, 83:-00644
17. Kim H, Park S, Lee T, Nahm B, Chung Y, Seo K, Kim HJJofp: Identification of Salmonella enterica serovar Typhimurium using specific PCR primers obtained by comparative genomics in Salmonella serovars. J Food Prot 2006, 69:1653-1661
18. Siala M, Barbana A, Smaoui S, Hachicha S, Marouane C, Kammoun S, Gdoura R, Messadi-Akrout F: Screening and Detecting Salmonella in Different Food Matrices in Southern Tunisia Using a Combined Enrichment/Real-Time PCR

- Method: Correlation with Conventional Culture Method. *Front Microbiol* 2017, 8:2416
19. Wang Y, Wang Y, Ma A-J, Li D-X, Luo L-J, Liu D-X, Jin D, Liu K, Ye C-Y: Rapid and sensitive isothermal detection of nucleic-acid sequence by multiple cross displacement amplification. *Sci Rep* 2015, 5:11902
  20. Wang Y, Wang Y, Zhang L, Liu D, Luo L, Li H, Cao X, Liu K, Xu J, Ye C: Multiplex, rapid, and sensitive isothermal detection of nucleic-acid sequence by endonuclease restriction-mediated real-time multiple cross displacement amplification. *Front Microbiol* 2016, 7:753
  21. Zhang X, Payne M, Lan R: In silico Identification of Serovar-Specific Genes for Salmonella Serotyping. *Front Microbiol* 2019, 10:835
  22. Beltran P, Plock SA, Smith NH, Whittam TS, Old DC, Selander RK: Reference collection of strains of the Salmonella typhimurium complex from natural populations. *J Gen Microbiol* 1991, 137:601-606
  23. Boyd EF, Wang F-S, Beltran P, Plock SA, Nelson K, Selander RK: Salmonella reference collection B (SARB): strains of 37 serovars of subspecies I. *J Gen Microbiol* 1993, 139:1125-1132
  24. Pang S, Octavia S, Feng L, Liu B, Reeves PR, Lan R, Wang LJBg: Genomic diversity and adaptation of Salmonella enterica serovar Typhimurium from analysis of six genomes of different phage types. *BMC Genomics* 2013, 14:718
  25. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD: SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* 2012, 19:455-477
  26. Yoshida CE, Kruczkiewicz P, Laing CR, Lingohr EJ, Gannon VP, Nash JH, Taboada EN: The Salmonella in silico typing resource (SISTR): an open web-accessible tool for rapidly typing and subtyping draft Salmonella genome assemblies. *PLoS One* 2016, 11: -0147101
  27. Zhang S, Yin Y, Jones MB, Zhang Z, Kaiser BLD, Dinsmore BA, Fitzgerald C, Fields PI, Deng X: Salmonella serotype determination utilizing high-throughput genome sequencing data. *J Clin Microbiol* 2015, 53:1685-1692

28. Treangen TJ, Ondov BD, Koren S, Phillippy AM: The Harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. *Genome Biol* 2014, 15:524
29. Schneider S, Roessli D, Excoffier LJUmV: Arlequin: a software for population genetics data analysis. *Evol Bioinform Online* 2000, 2: 2496-2497
30. Kumar S, Stecher G, Li M, Knyaz C, Tamura K: MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Mol Biol Evol* 2018, 35:1547-1549
31. Akiba M, Kusumoto M, Iwata TJJomm: Rapid identification of *Salmonella enterica* serovars, typhimurium, choleraesuis, infantis, hadar, enteritidis, dublin and gallinarum, by multiplex PCR. *J Microbiol Methods* 2011, 85:9-15
32. Bugarel M, Tudor A, Loneragan GH, Nightingale KKJJomm: Molecular detection assay of five *Salmonella* serotypes of public interest: Typhimurium, Enteritidis, Newport, Heidelberg, and Hadar. *J Microbiol Methods* 2017, 134:14-20
33. Kim S, Frye JG, Hu J, Fedorka-Cray PJ, Gautom R, Boyle DS: Multiplex PCR-based method for identification of common clinical serotypes of *Salmonella enterica* subsp. *enterica*. *J Clin Microbiol* 2006, 44:3608-3615
34. Park SH, Kim HJ, Cho WH, Kim JH, Oh MH, Kim SH, Lee BK, Ricke SC, Kim HYJFml: Identification of *Salmonella enterica* subspecies I, *Salmonella enterica* serovars Typhimurium, Enteritidis and Typhi using multiplex PCR. *FEMS Microbiol Lett* 2009, 301:137-146
35. Rodríguez-Lázaro D, Hernández M, Esteve T, Hoorfar J, Pla MJJoMM: A rapid and direct real time PCR-based method for identification of *Salmonella* spp. *J Microbiol Methods* 2003, 54:381-390
36. Shanmugasundaram M, Radhika M, Murali H, Batra HJWJoM, Biotechnology: Detection of *Salmonella enterica* serovar Typhimurium by selective amplification of *fliC*, *fljB*, *iroB*, *invA*, *rfbJ*, STM2755, STM4497 genes by polymerase chain reaction in a monoplex and multiplex format. *World J Microbiol Biotechnol* 2009, 25: 1385-1394
37. Xiong D, Song L, Pan Z, Jiao X: Identification and Discrimination of *Salmonella enterica* Serovar Gallinarum Biovars Pullorum and Gallinarum Based on a One-Step Multiplex PCR Assay. *Front Microbiol* 2018, 9:1718

38. Xiong D, Song L, Tao J, Zheng H, Zhou Z, Geng S, Pan Z, Jiao X: An efficient multiplex PCR-based assay as a novel tool for accurate inter-serovar discrimination of *Salmonella* Enteritidis, *S. Pullorum/Gallinarum* and *S. Dublin*. *Front Microbiol* 2017, 8:420
39. Ohtsuka K, Yanagawa K, Takatori K, Hara-Kudo YJAE: Detection of *Salmonella enterica* in naturally contaminated liquid eggs by loop-mediated isothermal amplification, and characterization of *Salmonella* isolates. *Appl Environ Microbiol* 2005, 71:6730-6735
40. Okamura M, Ohba Y, Kikuchi S, Suzuki A, Tachizaki H, Takehara K, Ikeda M, Kojima T, Nakamura MJV: Loop-mediated isothermal amplification for the rapid, sensitive, and specific detection of the O9 group of *Salmonella* in chickens. *Vet Microbiol* 2008, 132: 197-204
41. Ziros PG, Kokkinos PA, Papanotas K, Vantarakis AJTJoM, Biotechnology, Sciences F: Loop-mediated isothermal amplification (LAMP) for the detection of *Salmonella* spp. isolated from different food types. *J Microbiol Biotechnol Food Sci* 2012, 2:152-161
42. Gong J, Zhuang L, Zhu C, Shi S, Zhang D, Zhang L, Yu Y, Dou X, Xu B, Wang CJFp, disease: Loop-mediated isothermal amplification of the *sefA* gene for rapid detection of *Salmonella* Enteritidis and *Salmonella* Gallinarum in chickens. *Foodborne Pathog Dis* 2016, 13:177-181
43. Okamura M, Ohba Y, Kikuchi S, Takehara K, Ikeda M, Kojima T, Nakamura MJAd: Rapid, sensitive, and specific detection of the O4 group of *Salmonella enterica* by loop-mediated isothermal amplification. *Avian Dis* 2009, 53:216-221
44. Pavan Kumar P, Agarwal R, Thomas P, Sailo B, Prasannavadhana A, Kumar A, Kataria J, Singh DJFb: Rapid detection of *Salmonella enterica* subspecies *enterica* serovar Typhimurium by loop mediated isothermal amplification (LAMP) test from field chicken meat samples. *Food Biotechnol* 2014, 28:50-62
45. Ravan H, Yazdanparast RJWJoM, Biotechnology: Development of a new loop-mediated isothermal amplification assay for *prt* (*rfbS*) gene to improve the identification of *Salmonella* serogroup D. *World J Microbiol Bio-technol* 2012, 28:2101-2106
46. Yang JL, Yang R, Yang SQ, Fu LZ, Cheng AC, Wang MS, Zhang SH, Shen KF, Jia RY, Deng SXJJoam: Simple and rapid detection of *Salmonella* serovar

- Enteritidis under field conditions by loop-mediated isothermal amplification. *J Appl Microbiol* 2010, 109: 1715-1723
47. Yang Q, Domesle KJ, Ge BJFp, disease: Loop-mediated isothermal amplification for *Salmonella* detection in food and feed: current applications and future directions. *Foodborne Pathog Dis* 2018, 15:309-331
  48. Chen Z, Zhang K, Yin H, Li Q, Wang L, Liu ZJFS, Wellness H: Detection of *Salmonella* and several common *Salmonella* serotypes in food by loop-mediated isothermal amplification method. *Food Sci Human Wellness* 2015, 4:75-79
  49. Kardos G, Farkas T, Antal M, Nogrady N, Kiss IJLiam: Novel PCR assay for identification of *Salmonella enterica* serovar *Infantis*. *Lett Appl Microbiol* 2007, 45:421-425
  50. Chiang Y-C, Wang H-H, Ramireddy L, Chen H-Y, Shih C-M, Lin C-K, Tsen H-YJjof, analysis d: Designing a biochip following multiplex polymerase chain reaction for the detection of *Salmonella* serovars *Typhimurium*, *Enteritidis*, *Infantis*, *Hadar*, and *Virchow* in poultry products. *J Food Drug Anal* 2018, 26:58-66
  51. Deng X, Desai PT, den Bakker HC, Mikoleit M, Tolar B, Trees E, Hendriksen RS, Frye JG, Porwollik S, Weimer BC: Genomic epidemiology of *Salmonella enterica* serotype *Enteritidis* based on population structure of prevalent lineages. *Emerg Infect Dis* 2014, 20: 1481-1489
  52. Zou Q-H, Li R-Q, Liu G-R, Liu S-L: Genotyping of *Salmonella* with lineage-specific genes: correlation with serotyping. *Int J Infect Dis* 2016, 49:134-140
  53. Simpson KM, Hill-Cawthorne GA, Ward MP, Mor SMJBid: Diversity of *Salmonella* serotypes from humans, food, domestic animals and wildlife in New South Wales, Australia. *BMC Infect Dis* 2018, 18:623
  54. Selander RK, Smith N, Li J, Beltran P, Ferris K, Kopecko D, Rubin F: Molecular evolutionary genetics of the cattle-adapted serovar *Salmonella dublin*. *J Bacteriol* 1992, 174:3587-3592
  55. Jones MA, Wigley P, Page KL, Hulme SD, Barrow PA: *Salmonella enterica* serovar *Gallinarum* requires the *Salmonella* pathogenicity island 2 type III secretion system but not the *Salmonella* pathogenicity island 1 type III secretion system for virulence in chickens. *Infect Immun* 2001, 69:5471-5476

56. Kang MS, Kwon YK, Jung BY, Kim A, Lee KM, An BK, Song EA, Kwon JH, Chung GS: Differential identification of *Salmonella enterica* subsp. *enterica* serovar *Gallinarum* biovars *Gallinarum* and *Pullorum* based on polymorphic regions of *glgC* and *speC* genes. *Vet Microbiol* 2011, 147:181-185
57. Pal S, Dey S, Batabyal K, Banerjee A, Joardar SN, Samanta I, Isore DP: Characterization of *Salmonella Gallinarum* isolates from backyard poultry by polymerase chain reaction detection of invasion (*invA*) and *Salmonella* plasmid virulence (*spvC*) genes. *Vet World* 2017, 10:814-817
58. Uzzau S, Brown DJ, Wallis T, Rubino S, Leori G, Bernard S, Casadesús J, Platt DJ, Olsen JE, Infection: Host adapted serotypes of *Salmonella enterica*. *Epidemiol Infect* 2000, 125:229-255



# **Chapter 4. Cluster-specific gene markers enhance *Shigella* and Enteroinvasive *Escherichia coli* *in silico* serotyping**

## **4.1 Link to Thesis**

*Shigella* share ancestry within *E. coli* as well as similar physiological, biochemical and genetic characteristics with enteroinvasive *Escherichia coli* (EIEC). Misidentification of EIEC as *shigella* is common and crucially, distinguishing them is important for clinical, epidemiological and diagnostic investigations. As presented in Chapter 1, current genetic markers and *in silico* pipeline may not discriminate between *Shigella* and EIEC in all cases. Importantly, *Shigella* and EIEC are separated into multiple phylogenetic clusters. I therefore took advantage of the large number of the genome sequences for *Shigella* and EIEC in public databases to enhance the molecular identification and differentiation of *Shigella* and EIEC using specific genomic markers. Given the relatively poor performance of existing tools, I developed an *in silico* program using these markers that is capable of highly accurate molecular characterization of *Shigella* and EIEC. This chapter addresses the third aim of this thesis.

I have submitted this work to Microbial genomics 04/02/2021:

Zhang X, Payne M, Nguyen T, Kaur S, Lan R. Cluster-specific gene markers enhance *Shigella* and Enteroinvasive *Escherichia coli* *in silico* serotyping.

I have presented this work at national conference:

Zhang X, Payne M, Nguyen T, Kaur S, Lan R. Cluster-specific gene markers enhance *Shigella* and Enteroinvasive *Escherichia coli* *in silico* serotyping. Poster presentation, Australian Society for Microbiology Annual Scientific Meeting 2021.

**The Supplementary Material for this article can be found online at:**

<https://drive.google.com/drive/folders/1kEhrqzKOSWBr3ldvUvDpp9J4Y1KoMz0?usp=sharing>

**Supplemental material for this article is also listed at Appendix III.**

## 4.2 Abstract

*Shigella* and enteroinvasive *Escherichia coli* (EIEC) cause human bacillary dysentery with similar invasion mechanisms and share similar physiological, biochemical and genetic characteristics. The ability to differentiate *Shigella* and EIEC from each other is important for clinical diagnostic and epidemiologic investigations. The existing genetic signatures may not discriminate between *Shigella* and EIEC. However, phylogenetically, *Shigella* and EIEC strains are composed of multiple clusters and are different forms of *E. coli*. In this study, we identified 10 *Shigella* clusters, 7 EIEC clusters and 53 sporadic types of EIEC by examining over 17,000 publicly available *Shigella* and EIEC genomes. We compared *Shigella* and EIEC accessory genomes to identify the cluster-specific gene marker sets for the 17 clusters and 53 sporadic types. The gene marker sets showed 99.64% accuracy and more than 97.02% specificity. In addition, we developed a freely available *in silico* serotyping pipeline named *Shigella* EIEC Cluster Enhanced Serotype Finder (ShigEiFinder) by incorporating the cluster-specific gene markers and established *Shigella* and EIEC serotype specific O antigen genes and modification genes into typing. ShigEiFinder can process either paired end Illumina sequencing reads or assembled genomes and almost perfectly differentiated *Shigella* from EIEC with 99.70% and 99.74% cluster assignment accuracy for the assembled genomes and mapped reads respectively. ShigEiFinder was able to serotype over 59 *Shigella* serotypes and 22 EIEC serotypes and provided a high specificity with 99.40% for assembled genomes and 99.38% for mapped reads for serotyping. The cluster-specific gene markers and our new serotyping tool, ShigEiFinder (installable package: <https://github.com/LanLab/ShigEiFinder>, online tool: <https://mgtdb.unsw.edu.au/ShigEiFinder/>), will be useful for epidemiologic and diagnostic investigations.

## 4.3 Introduction

*Shigella* is a leading cause of diarrhea with a very low infective dose (1, 2). The infections can vary from mild diarrhea to severe bloody diarrhea referred to as bacillary dysentery.

The estimated cases of *Shigella* infections are 190 million with at least 210,000 deaths annually, predominantly in children younger than 5 years old in developing countries (3-7). *Shigella* infections also have a significant impact on public health in developed countries, although most cases are travel-associated (8).

The *Shigella* genus consists of four species, *Shigella sonnei*, *Shigella flexneri*, *Shigella boydii* and *Shigella dysenteriae* (9). Serological testing further classifies *Shigella* species into more than 55 serotypes through the agglutination reaction of antisera to *Shigella* serotype specific O-antigens (10, 11). Up to 89.6% *Shigella* infections were caused by *S. flexneri* (65.9%) and *S. sonnei* (23.7%) globally (12, 13). The predominant serotype reported in *Shigella* infections has been *S. flexneri* serotype 2a while *S. dysenteriae* serotype 1 has caused the most severe disease (10, 14). Note that for brevity, in all references to *Shigella* serotypes below, *S. sonnei*, *S. flexneri*, *S. boydii* and *S. dysenteriae* are abbreviated as SS, SF, SB and SD respectively and a serotype is designated with an abbreviated “species” name plus the serotype number e.g. *S. dysenteriae* serotype 1 is abbreviated as SD1.

Enteroinvasive *Escherichia coli* (EIEC) is a pathovar of *E. coli* that causes diarrhoea with less severe symptoms than *Shigella* infections in humans worldwide, particularly in developing countries (8, 13, 15-18). EIEC infections in developed countries are mainly imported (19). EIEC has more than 18 specific *E. coli* O-serotypes (19, 20). Although the incidence of EIEC is low (17), EIEC serotypes have been associated with outbreaks and sporadic cases of infections (20-22). In contrast to *Shigella*, EIEC infections are not notifiable in many countries (23, 24).

*Shigella* and EIEC have always been considered very closely related and share several characteristics (25-28). *Shigella* and EIEC are both non-motile and lack the ability of fermenting lactose (24). Some EIEC O antigens are identical or similar to *Shigella* O antigens (O112ac, O124, O136, O143, O152 and O164) (26, 29-31). Furthermore, *Shigella* and EIEC both carry the virulence plasmid pINV, which encodes virulence genes required for invasion (32, 33) and contain *ipaH* (invasion plasmid antigen H) genes with the exception of some SB13 isolates (11, 23, 24, 34, 35). *Shigella* and EIEC have arisen from *E. coli* in multiple independent events and should be regarded as a single pathovar

of *E. coli* (25, 26, 28, 36-38). Previous phylogenetic studies suggested that *Shigella* isolates were divided into 3 clusters (C1, C2 and C3) with 5 outliers (SS, SB13, SD1, SD8 and SD10) (25, 28) whereas EIEC isolates were grouped into four clusters (C4, C5, C6 and C7) (26). The seven *Shigella* and EIEC clusters and 5 outliers of *Shigella* are within the broader non-enteroinvasive *E. coli* species except for SB13 which is closer to *Escherichia albertii* (39, 40). WGS-based phylogenomic studies have also defined multiple alternative clusters of *Shigella* and EIEC (23, 28, 41).

The traditional biochemical test for motility and lysine decarboxylase (LDC) activity (42) and molecular test for the presence of *ipaH* gene have been used to differentiate *Shigella* and EIEC from non-enteroinvasive *E. coli* (24, 43-45). Agglutination with *Shigella* and EIEC associated antiserum further classifies *Shigella* or EIEC to serotype level. However, cross-reactivity, strains not producing O antigens, and newly emerged *Shigella* serotypes may all prevent accurate serotyping (11, 46). Serotyping by antigenic agglutination is being replaced by molecular serotyping (46-48), which can be achieved through examination of the sequences of O antigen biosynthesis and modification genes (8, 24, 49-52).

Recently, PCR-based molecular detection methods targeting the gene *lacY* were developed to distinguish *Shigella* from EIEC (53, 54). However, the ability of the primers described in these methods to accurately differentiate between *Shigella* and EIEC was later questioned (23, 28). With the uptake of whole-genome sequencing technology, several studies have identified phylogenetic clade specific markers, species specific markers and EIEC lineage-specific genes for discrimination between *Shigella* and EIEC and between *Shigella* species (23, 27, 28, 41, 55, 56). More recently, genetic markers *lacY*, *cadA*, *Ss\_methylase* were used for identification of *Shigella* and EIEC (11). However, these markers failed to discriminate between *Shigella* and EIEC when a larger genetic diversity is considered (23, 28, 55). A *Kmer*-based approach can identify *Shigella* isolates to the species level but misidentification was also observed (56).

In this study, we aimed to i), identify phylogenetic clusters of *Shigella* and EIEC through large scale examination of publicly available genomes; ii), identify cluster-specific gene markers using comparative genomic analysis of *Shigella* and EIEC accessory genomes

for differentiation of *Shigella* and EIEC; iii), develop a pipeline for *Shigella* and EIEC *in silico* serotyping based on the cluster-specific gene markers combined with *Shigella* and EIEC serotype-specific O antigen and H antigen genes. We demonstrate that these cluster-specific gene markers enhance *in silico* serotyping using genomic data. We also developed an automated pipeline for cluster typing and serotyping of *Shigella* and EIEC from WGS data.

## 4.4 Materials and Methods

### 4.4.1 Identification of *Shigella* and EIEC isolates from NCBI database

*E. coli* and *Shigella* isolates from the NCBI SRA (National Center for Biotechnology Information Sequence Read Archive) in May of 2019 were queried. The keywords “*Escherichia coli*” and “*Shigella*” were used to retrieve SRA accession numbers of *E. coli* and *Shigella* isolates. Raw reads were retrieved from the ENA (European Nucleotide Archive). The *ipaH* gene (GenBank accession number M32063.1) was used to screen *E. coli* and *Shigella* reads using Salmon v0.13.0 (57). Taxonomic classification for *E. coli* and *Shigella* was confirmed by Kraken v1.1.1 (58). Molecular serotype prediction of *ipaH* negative *Shigella* isolates was performed by ShigaTyper v1.0.6 (11). Isolates that were *ipaH* positive and isolates with designation of SB13 by ShigaTyper were selected to form the *Shigella* and EIEC database.

The sequence types (STs) and ribosomal STs (rSTs) of *ipaH* gene negative *E. coli* (non-enteroinvasive *E. coli*) isolates were examined. STs and rSTs for these isolates were obtained from the *E. coli* and *Shigella* database in Enterobase (59) in May of 2019. For STs and rSTs with only one isolate, the isolates were selected. For STs and rSTs with more than one isolate, one representative isolate for each ST and rST were randomly selected. In total, 12,743 *ipaH* negative *E. coli* isolates representing 3,800 STs and 11,463 rSTs were selected as a non-enteroinvasive *E. coli* control database.

### 4.4.2 Genome sequencing

Whole-genome sequencing (WGS) of 31 EIEC strains used in a previous study (26) was performed by Illumina NextSeq (Illumina, Scoresby, VIC, Australia). DNA libraries were constructed using Nextera XT Sample preparation kits (Illumina Inc., San Diego, CA, USA) and sequenced using the NextSeq sequencer (Illumina Inc.). FASTQ sequences of

the strains sequenced in this study were deposited in the NCBI under the BioProject (PRJNA692536).

#### **4.4.3 Genome assembly and data processing**

Raw reads were *de novo* assembled using SPADes v3.14.0 assembler with default settings [<http://bioinf.spbau.ru/spades>] (60). The metrics of assembled genomes were obtained with QUAST v5.0.0 (61). Three standard deviations (SD) from the mean for contig number, largest contig, total length, GC, N50 and genes were used as quality filters for assembled genomes.

The STs for isolates in the *Shigella* and EIEC database were checked by using mlst (<https://github.com/tseemann/mlst>) with the *E. coli* scheme from PubMLST (62). rSTs were extracted from the *E. coli* and *Shigella* rMLST database in Enterobase (59) in May of 2019. Serotype prediction for isolates in *Shigella* and EIEC database was performed by ShigaTyper v1.0.6 (11). Serotyping of *E. coli* O and H antigens were predicted by using SerotypeFinder v2.0.1 (63).

#### **4.4.4 Selection of isolates for *Shigella* and EIEC identification dataset**

The selection of isolates for the identification dataset was based on the representative isolates for each ST, rST and serotype of *Shigella* and EIEC in the *Shigella* and EIEC database. For STs, rSTs and serotypes with only one isolate, the isolate was selected. For STs, rSTs and serotypes with more than one isolate, one representative isolate for each ST, rST and serotype was randomly selected. 72 ECOR isolates downloaded from Enterobase (59) and 18 *E. albertii* isolates were used as controls for the identification dataset. The details of the identification dataset are listed in Table S1. The remaining isolates in the *Shigella* and EIEC database were referred as the validation dataset (Table S2).

The identification dataset was used to characterise the phylogenetic relationships of *Shigella* and EIEC. The identification dataset was also used to identify cluster-specific gene markers. The validation dataset was used to evaluate the performance of cluster-specific gene markers using the *in silico* serotyping pipeline.

#### **4.4.5 Phylogeny of *Shigella* and EIEC based on WGS**

Nine phylogenetic trees including an identification tree, a confirmation tree and 7 validation trees were constructed using Quiktree v1.3 (64) with the default parameters to identify and confirm the phylogenetic clustering of *Shigella* and EIEC isolates. The phylogenetic trees were visualised by Grapetree and ITOL v5 (65, 66).

The identification phylogenetic tree was generated based on isolates in the identification dataset for the characterisation of clusters of *Shigella* and EIEC isolates (Fig. 1). A subset of 485 isolates known to represent each identified cluster from the identification dataset were then selected. The confirmation tree was constructed based on the subset of 485 isolates from the identification dataset and 1,872 non-enteroinvasive *E. coli* isolates from non-enteroinvasive *E. coli* control dataset (2,357 isolates total). This tree was used for confirmation of the phylogenetic relationships between identified *Shigella* and EIEC clusters in the identification dataset and non-enteroinvasive *E. coli* isolates. The validation trees were generated based on *Shigella* and EIEC isolates from the validation dataset and a subset of 575 isolates from the identification dataset to assign validation dataset isolates to the clusters defined.

#### **4.4.6 Investigation of *Shigella* virulence plasmid pINV**

The presence of *Shigella* virulence plasmid pINV in isolates were investigated by using BWA-MEM v0.7.17 (Burrows-Wheeler Aligner) (67) to align isolate raw reads onto the reference sequence of pINV (68) (NC\_024996.1). Mapped reads were sorted and indexed using Samtools v1.9 (69). The individual gene coverage from mapping was obtained using Bedtools coverage v2.27.1 (70).

#### **4.4.7 Identification of the cluster-specific gene markers**

Cluster-specific gene markers were identified from *Shigella* and EIEC accessory genomes. The genomes from the identification dataset were annotated using PROKKA v1.13.3 (71). Pan- and core-genomes were analysed using roary v3.12.0 (72) using an 80% sequence identity threshold. An in-house python script was used to generate the candidate specific gene markers for each cluster from the profile of gene presence or absence in each genome which was produced by roary. The script is available on <https://github.com/LanLab/ShigEiFinder/tree/main/scripts> and the process to identify

potential candidates is described in Data S1. The best performance cluster-specific gene marker set was selected from the candidates by using BLASTN to search against the identification dataset.

In this study, the genomes from a given cluster containing all specific gene markers for that cluster were termed true positives (TP), the genomes from the same cluster lacking any of those same gene markers were termed false negatives (FN). The genomes from other clusters containing all of those same gene markers were termed false positives (FP). Relaxed cut-offs (40% FP) were used in initial screening to ensure that all clusters had candidate specific gene markers which could be further investigated.

The sensitivity (True positive rate, TPR) of each cluster-specific gene marker was defined as  $TP/(TP+FN)$ . The specificity (True negative rate, TNR) was defined as  $TN/(TN+FP)$ .

#### **4.4.8 Validation of the cluster-specific gene markers**

The ability of cluster-specific gene markers to assign *Shigella* and EIEC isolates was examined by using BLASTN to search against the validation dataset (Table S2) and non-enteroinvasive *E. coli* control database for the presence of any of the cluster-specific gene marker set. The BLASTN thresholds were defined as 80% sequence identity and 50% gene length coverage.

#### **4.4.9 Development of ShigEiFinder, an automated pipeline for molecular serotyping of *Shigella* and EIEC**

ShigEiFinder was developed using paired end illumina genome sequencing reads or assembled genomes to type *Shigella* and EIEC isolates to serotype level using cluster-specific gene markers combined with *Shigella* and EIEC serotype specific O antigen genes (*wzx* and *wzy*) and modification genes (Fig. 2). Further details of the algorithms used were presented in Data S2. We used the same signature O and H sequences from ShigaTyper and SerotypeFinder (Data S3) (11, 63). These include *Shigella* serotype-specific *wzx/wzy* genes and modification genes from ShigaTyper and *E. coli* O antigen and *fliC* (H antigen) genes from SerotypeFinder. *ipaH* gene and 38 virulence genes used in analysis of virulence of 59 sporadic EIEC isolates were also included in the typing



reference sequences database. Seven House Keeping (HK) genes -*recA*, *purA*, *mdh*, *icd*, *gyrB*, *fumC* and *adk* downloaded from NCBI were used for contamination checking.

For raw reads input, raw reads were aligned to the typing reference sequences by using BWA-MEM v0.7.17 (67). The mapping length percentage and the mean mapping depth for all genes were calculated using Samtools coverage v1.10 (69). To determine whether the genes were present or absent, 50% of mapping length for all cluster-specific gene markers, virulence genes, O antigen genes and 10% for *ipaH* gene were used as cutoff value. The ratio of mean mapping depth to the mean mapping depth of the 7 HK genes was used to determine a contamination threshold with ratios less than 1% for *ipaH* gene and less than 10% for other genes assigned as contamination. Reads coverage mapped to particular regions of genes were checked by using samtools mpileup v1.10 (69).

For assembled genome input, assembled genomes were searched against the typing reference sequences using BLASTN v2.9.0 (73) with 80% sequence identity and 50% gene length coverage for all genes with exception of *ipaH* gene which was defined as 10% gene length coverage.

ShigEiFinder was tested with the identification dataset and validated with the *Shigella* and EIEC validation dataset and non-enteroinvasive *E. coli* control database. The specificity defined as  $(1 - \frac{\text{the number of non-enteroinvasive } E. coli \text{ isolates being detected}}{\text{the total number of non-enteroinvasive } E. coli \text{ isolates}}) * 100$ .

## 4.5 Results

### 4.5.1 Screening sequenced genomes for *Shigella* and EIEC isolates

We first screened available *E. coli* and *Shigella* genomes based on the presence of the *ipaH* gene. We examined 122,361 isolates with the species annotation of *E. coli* (104,256) or *Shigella* (18,105) with paired-end Illumina sequencing reads available in NCBI SRA database. Of 122,361 isolates, 17,989 isolates were positive to the *ipaH* gene including 455 out of 104,256 *E. coli* isolates and 17,434 out of 18,105 *Shigella* isolates. The 17,989 *ipaH* positive *E. coli* and *Shigella* isolates and 571 *ipaH* negative “*Shigella*” isolates were checked for taxonomic classification and genome assembly quality using the methods described in the Materials and Methods. 17,320 *ipaH* positive *E. coli* and *Shigella*

genomes and 246 *ipaH* negative “*Shigella*” genomes passed quality filters. Among 246 *ipaH* negative “*Shigella*” isolates, 11 isolates were predicted as SB13 by using ShigaTyper (11) while the remaining 235 isolates were classified with taxonomic identifier of *E. coli* by Kraken v1.1.1 (58) and their *E. coli* O/H antigen types predicted using SerotypeFinder were not classic EIEC serotypes or their O antigen untypable. These 235 isolates were removed from analysis. A total of 17,331 isolates including 17,320 *ipaH* positives and 11 SB13 isolates were selected to form the *Shigella* and EIEC database. The *Shigella* and EIEC database contained 429 isolates with species identifier of *E. coli* and 16,902 isolates with species identifier of *Shigella*.

Isolates in the *Shigella* and EIEC database were typed using MLST, ShigaTyper and SerotypeFinder. MLST and rMLST divided the 17,331 *Shigella* and EIEC isolates into 252 STs (73 isolates untypeable by MLST) and 1,128 rSTs (3,513 isolates untypeable by rMLST). Of 16,902 isolates with species identifier of *Shigella*, 8,313 isolates and 8,189 isolates were typed as *Shigella* and EIEC respectively by ShigaTyper while 400 isolates were untypeable. ShigaTyper typed the majority of the 8,313 isolates as SF (66.82%) including 25.43% SF2a isolates, followed by SS (19.69%), SB (7.22%) and SD (6.27%).

SerotypeFinder typed 293 of the 429 *E. coli* isolates into 71 *E. coli* O/H antigen types. Among these 293 isolates with typable O/H antigen types, 190 isolates belonged to 22 known EIEC serotypes (O28ac:H-, O28ac:H7, O29:H4, O112ac:H26, O121:H30, O124:H30, O124:H24, O124:H7, O132:H7, O132:H21, O135:H30, O136:H7, O143:H26, O144:H25, O152:H-, O152:H30, O164:H-, O164:H30, O167:H26, O173:H7 and 2 newly emerged EIEC serotypes O96:H19 and O8:H19) (20-22). The remaining 136 of the 429 isolates were O antigen untypable and typed to 15 H antigen types only by SerotypeFinder, of which H16 was the predominant type.

#### **4.5.2 Identification of *Shigella* and EIEC clusters**

*Shigella* and EIEC are known to have been derived from *E. coli* independently. To identify previously defined clusters (25, 26) and any new clusters from the 17,331 *Shigella* and EIEC isolates, we selected representative isolates to perform phylogenetic analysis as it was impractical to construct a tree with all isolates. The selection was based on ST, rST and serotype of the 17,331 *Shigella* and EIEC isolates. One isolate was

selected to represent each ST, rST and serotype for a total of 1,830 isolates. Note that in the case that STs or rSTs overlapped with serotype, an isolate would have only selected once to avoid duplicates of the same isolate. The selection included 252 STs, 1,128 rSTs, 59 *Shigella* serotypes (21 SB serotypes, 20 SF serotypes, 17 SD serotypes and 1 SS serotype), 22 EIEC known serotypes and 31 other or partial antigen types. A further 31 in-house sequenced EIEC isolates, 18 EIEC isolates used in a previous typing study (41), 72 ECOR isolates and 18 *E. albertii* isolates were also included to form the identification dataset of 1,969 isolates. Details are listed in Table S1. A phylogenetic tree was constructed based on the identification dataset to identify the clusters (Fig. 1).

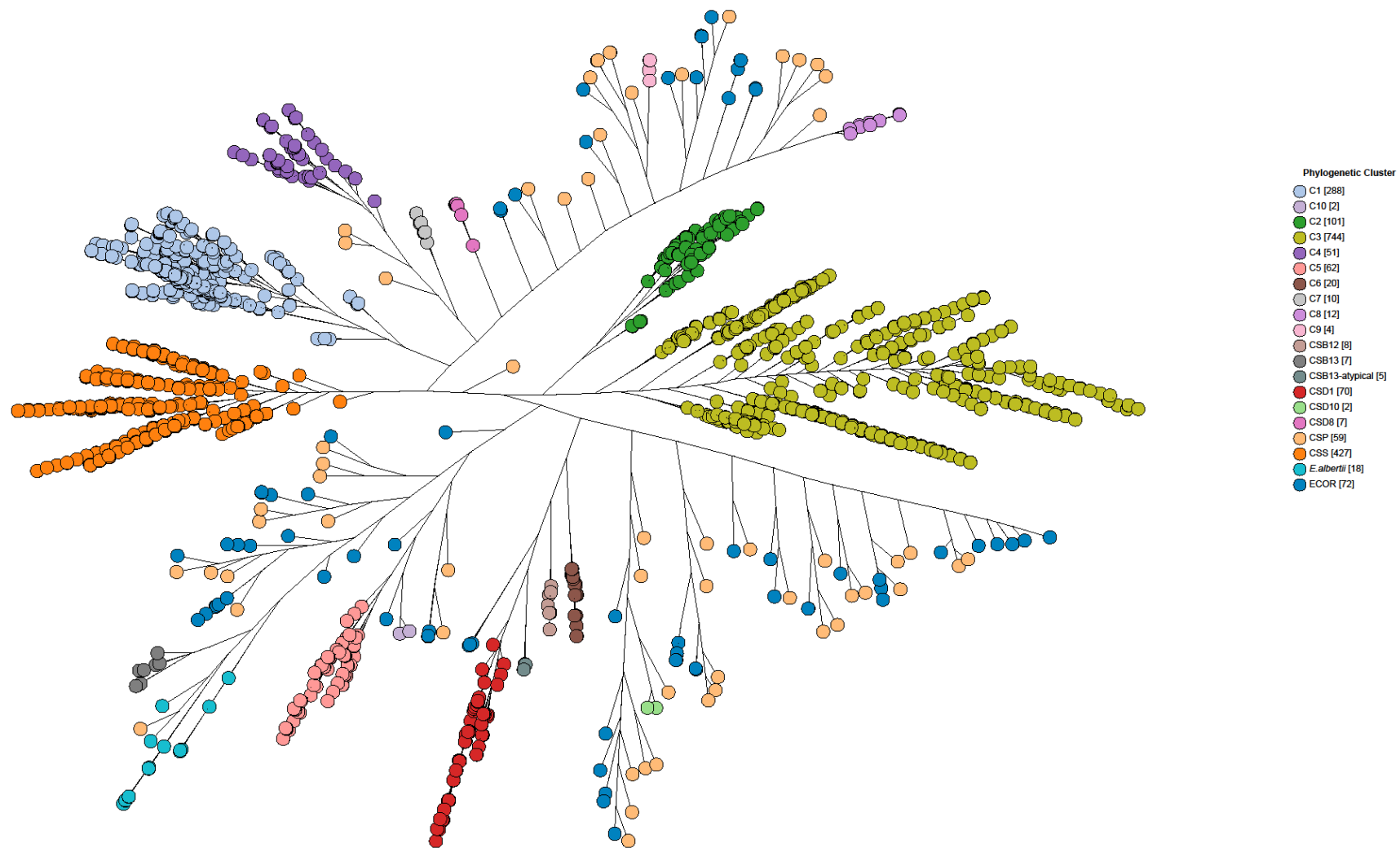
All known clusters were identified (Fig. 1) including 3 *Shigella* clusters (C1, C2, C3) and 5 outliers (SD1, SD8, SD10, SB13 and SS) as defined by Puppo et al (25) and 4 EIEC clusters (C4, C5, C6 and C7) defined by Lan et al [26](#). Each of these clusters was supported by a bootstrap value of 80% or greater (Fig. S1). 1,789 isolates of the 1,879 *Shigella* and EIEC isolates (1,830 isolates from the *Shigella* and EIEC database, 31 in-house sequenced EIEC isolates and 18 EIEC isolates from Hazen *et al.* (41)) fell within these clusters.

Of the remaining 90 *Shigella* and EIEC unclustered isolates, 31 belonged to typical or known *Shigella* or EIEC serotypes including 5 SB13 isolates, 8 SB12 isolates, 2 EIEC O135:H30 isolates, 12 EIEC O96:H19 isolates and 4 EIEC O8:H19 isolates, while 59 isolates were separated from the identified clusters by non-*Shigella*/EIEC isolates and interspersed among non-*Shigella*/EIEC isolates. Of the 59 isolates, 34 isolates were singletons with one isolate as sole member of the group while the remaining 25 isolates formed 12 groups of 2 or more isolates. Furthermore their *E. coli* O/H antigen types were not classic EIEC serotypes or their O antigen untypable. These 59 isolates were named as sporadic EIEC isolates which are described in detail in the separate section below.

The 5 SB13 isolates were grouped into one lineage within *E. coli* and close to known *Shigella* and EIEC clusters rather than the established SB13 cluster outside *E. coli* which was within the *E. albertii* species. The former was previously named as atypical SB13 while the latter was previously named as typical SB13 (39). The 8 SB12 isolates formed one single cluster close to SD1 and atypical SB13 clusters. SB12 was previously grouped

into C3 based on housekeeping gene trees (25, 28) but was seen as outliers in two other studies (28, 56). Two EIEC O135:H30 isolates were grouped as a separate cluster close to C5. Twelve isolates belonging to EIEC serotype O96:H19 and 4 isolates typed as O8:H19 were clustered into two separate clusters, both of which were more closely related to SD8 than other *Shigella* and EIEC clusters. Each of these 5 groups was phylogenetically distinct and represented the classic *Shigella* or EIEC serotypes. Furthermore, each of the 5 groups was supported by a bootstrap value of 80% or greater (Fig. S1). Therefore, atypical SB13 and SB12 were defined as new clusters of *Shigella* while EIEC O96:H19, EIEC O8:H19 and EIEC O135:H30 were defined as C8, C9 and C10 respectively. In total there were 10 *Shigella* clusters and 7 EIEC clusters (Table 1).

**Chapter 4. Figure 1: *Shigella* and EIEC cluster identification phylogenetic tree.** Representative isolates from the identification dataset were used to construct the phylogenetic tree by Quiktree v1.3 (64) to identify *Shigella* and EIEC clusters and visualised by Grapetree. The dendrogram tree shows the phylogenetic relationships of 1879 *Shigella* and EIEC isolates represented in the identification dataset. Branch lengths are log scale for clarity. The tree scales indicated the 0.2 substitutions per locus. *Shigella* and EIEC clusters are coloured. Numbers in square brackets indicate the number of isolates for each identified cluster. CSP is sporadic EIEC lineages.



0.2

**Chapter 4. Table 1: The summary of identified *Shigella* and EIEC clusters and outliers in identification dataset**

Clusters <sup>#</sup>	No of isolates	No. STs	No. rSTs	Serotypes
C1 (25)	288	36	166	SB1-4, SB6, SB8, SB10, SB14, SB18, SB11 <sup>b</sup> , SB19-20 <sup>b</sup> ; SD3-7, SD9, SD11-13, SD14-15 <sup>b</sup> , SD-96-26b <sup>b</sup> ; SF6
C2 (9)	101	19	56	SB5, SB7, SB9, SB11, SB15, SB16, SB17; SD2, SD-E670-74 <sup>b</sup>
C3 (20)	744	81	437	SF1a, SF1b, SF1c (7a), SF2a, SF2b, SF3a, SF3b, SF4a, SF4av, SF4b, SF4bv, SF5a, SF5b, SF7b, SFX, SFXv (4c), SFY, SFYv, SF novel serotype; SB-E1621-54 <sup>b</sup>
C4 (9)	51	6	21	O28ac:H7/H-, O136:H7, O164:H7/H-, O29:H4, O173:H7, O124:H7, O132:H7 <sup>b</sup>
C5 (6)	62	4	15	O121:H30, O124:H30, O164:H30, O132:H21, O152:H30/H-
C6 (3)	20	2	6	O143:H26, O167:H26, O112ac:H26 <sup>b</sup>
C7	10	1	3	O144:H25
C8 <sup>a</sup>	12	2	1	O96:H19
C9 <sup>a</sup>	4	1	2	O8:H19
C10 <sup>a</sup>	2	1	1	O135:H30
CSS	427	39	294	SS
CSD1	70	8	56	SD1
CSD8	7	3	3	SD8
CSD10	2	2	1	SD10
CSB12 <sup>a</sup>	8	2	6	SB12
CSB13	7	3	3	SB13
CSB13-atypical <sup>a</sup>	5	3	3	SB13
Sporadic EIEC lineages <sup>a</sup> (53)	59	49	53	53 antigen types

<sup>#</sup>: Numbers in parentheses are the number of serotypes within that cluster. <sup>a</sup>: Clusters identified as new clusters in this study. <sup>b</sup>: Serotypes were inconsistent with previous analyses.

#### 4.5.3 Analysis of the 59 sporadic EIEC isolates

To determine the phylogenetic relationships of the above defined clusters and the remaining 59 sporadic EIEC isolates within the larger non-enteroinvasive *E. coli* population a confirmation tree was generated using 485 isolates representing the known clusters and 1,872 representative non-*Shigella*/EIEC isolates (Fig. S2). The 59 sporadic EIEC isolates were interspersed among non-*Shigella*/EIEC isolates and did not form large clusters. Groups of these isolates that were not previously identified were named as sporadic EIEC lineages followed by their serotype. For example, isolate M2330 (O152:H51) we sequenced in this study was named ‘sporadic EIEC lineage O152:H51’. There were 53 sporadic EIEC lineages including 5 lineages with 2 or more isolates and 48 lineages with only one isolate. The STs, rSTs and antigen types of these 59 isolates were listed in Table S1.

Some of the sporadic EIEC isolates fell into STs containing *ipaH* negative isolates. We therefore examined the presence of the pINV virulence plasmid in the sporadic EIEC isolates. We selected 38 genes that are essential for virulence including 35 genes (12 *mxi* genes, 9 *spa* genes, 5 *ipaA-J* genes, 6 *ipgA-F* genes as well as *acp*, *virB*, *icsB*) in the conserved entry region encoding the Mxi-Spa-Ipa type III secretion system and its effectors and 3 regulator genes (*virF*, *virA* and *icsA/virG*) (24, 32, 68) and determined the presence of pINV in the 59 sporadic EIEC isolates by mapping the sequence reads onto a pINV reference sequence (68). Reads from 18 non-*Shigella*/EIEC isolates that shared the same ST as one of 59 sporadic isolates were also mapped onto a pINV reference sequence (68).

The number of essential virulence genes with mapped reads in the 59 sporadic EIEC isolates were analysed (Fig. S3). Those isolates containing more than 25 of the 38 essential virulence genes were defined as virulence plasmid positive. While isolates containing between 13 and 25 were defined as intermediate and less than 13 were defined as virulence plasmid negative.

The 2 newly sequenced sporadic EIEC isolates (M2330 and M2339) were positive for the virulence plasmid, and of the other 57 sporadic EIEC isolates, 39 were positive, 9 were negative and another 9 were intermediate (Table S1). The results were compared with 18

non-*Shigella*/EIEC isolates mentioned above. The virulence plasmid was absent in all non-*Shigella*/EIEC isolates while all sporadic EIEC isolates in these STs were either positive or intermediate. Therefore, this analysis confirmed the sporadic isolates belonged to EIEC and the STs contained both EIEC and non- EIEC isolates.

#### **4.5.4 Identification of cluster-specific gene markers**

In this study, cluster-specific gene marker sets (single gene or two or more genes) were either present in all isolates of a cluster and absent in all other isolates. For the marker sets with two or more genes, a subset of cluster-specific genes for a given cluster could be found in other clusters but the entire set was only found in the target cluster.

Comparative genomic analysis on 1,969 accessory genomes from the identification dataset was used to identify the potential cluster-specific gene marker sets. Multiple candidate cluster-specific gene marker sets for each of the 17 *Shigella* and EIEC clusters and 53 sporadic EIEC lineages were identified through initial screening of the accessory genes from the 1,969 genomes. Genes associated with *Shigella* and EIEC O antigen clusters were excluded from the analysis. The candidate cluster-specific gene marker sets were 100% sensitive to clusters but with varying specificity. The cluster-specific gene marker sets with the lowest FP rates were then selected from candidate cluster-specific gene marker sets by BLASTN searches against genomes in the identification dataset using 80% sequence identity and 50% gene length threshold.

The cluster-specific gene marker sets were all 100% sensitive and 100% specific with the exception of those for C1 (99.94% specificity), C3 (99.91% specificity) and SS (99.8% specificity). The sensitivity and specificity for each cluster-specific gene marker or marker set for the identification dataset were listed in Table 2. A single specific gene for each of the 53 sporadic EIEC lineages were also selected with the exception of sporadic EIEC lineage 27 which has a set of 2 genes. These genes were all 100% sensitive and specific for a given sporadic EIEC lineage.



**Chapter 4. Table 2: The sensitivity and specificity of cluster-specific genes**

Clusters	Cluster-specific genes (Single/sets)	Identification dataset (1,969 isolates)		
		No of isolates	Sensitivity	Specificity
C1	Set of 4 genes	288	100	99.94 <sup>a</sup>
C2	Set of 3 genes	101	100	100
C3	Set of 3 genes	744	100	99.59 <sup>a</sup>
C4	Set of 2 genes	51	100	100
C5	Set of 3 genes	62	100	100
C6	Set of 2 genes	20	100	100
C7	Single gene	10	100	100
C8	Set of 2 genes	12	100	100
C9	Set of 2 genes	4	100	100
C10	Single gene	2	100	100
CSS	Set of 5 genes	427	100	99.87 <sup>a</sup>
CSD1	Set of 2 genes	70	100	100
CSD8	Single gene	7	100	100
CSD10	Single gene	2	100	100
CSB12	Single gene	8	100	100
CSB13	Single gene	7	100	100
CSB13-atypical	Single gene	5	100	100
53 CSP	Single gene / lineage	59	100	100

<sup>a</sup>:The specificity of cluster-specific gene set less than 100% was due to at least one FP found in that set. CSP: Sporadic EIEC lineages

All 37 cluster-specific gene markers and 54 sporadic EIEC lineages specific gene markers were located on the chromosome except that one of the C4 gene markers and 5 sporadic EIEC lineages specific genes were located on plasmids by NCBI BLAST searches. None of the cluster-specific gene markers were contiguous in the genomes. The location of these cluster-specific gene markers was determined by BLASTN against representative complete genomes of *Shigella* and EIEC containing gene features downloaded from GenBank (Accession number were listed in Table S3). In those cluster or sporadic lineages with no representative complete genome, specific gene markers were named using their cluster or sporadic EIEC lineage followed by the cluster or lineage number. For example, C7 specific gene marker was named “C7 specific gene”.

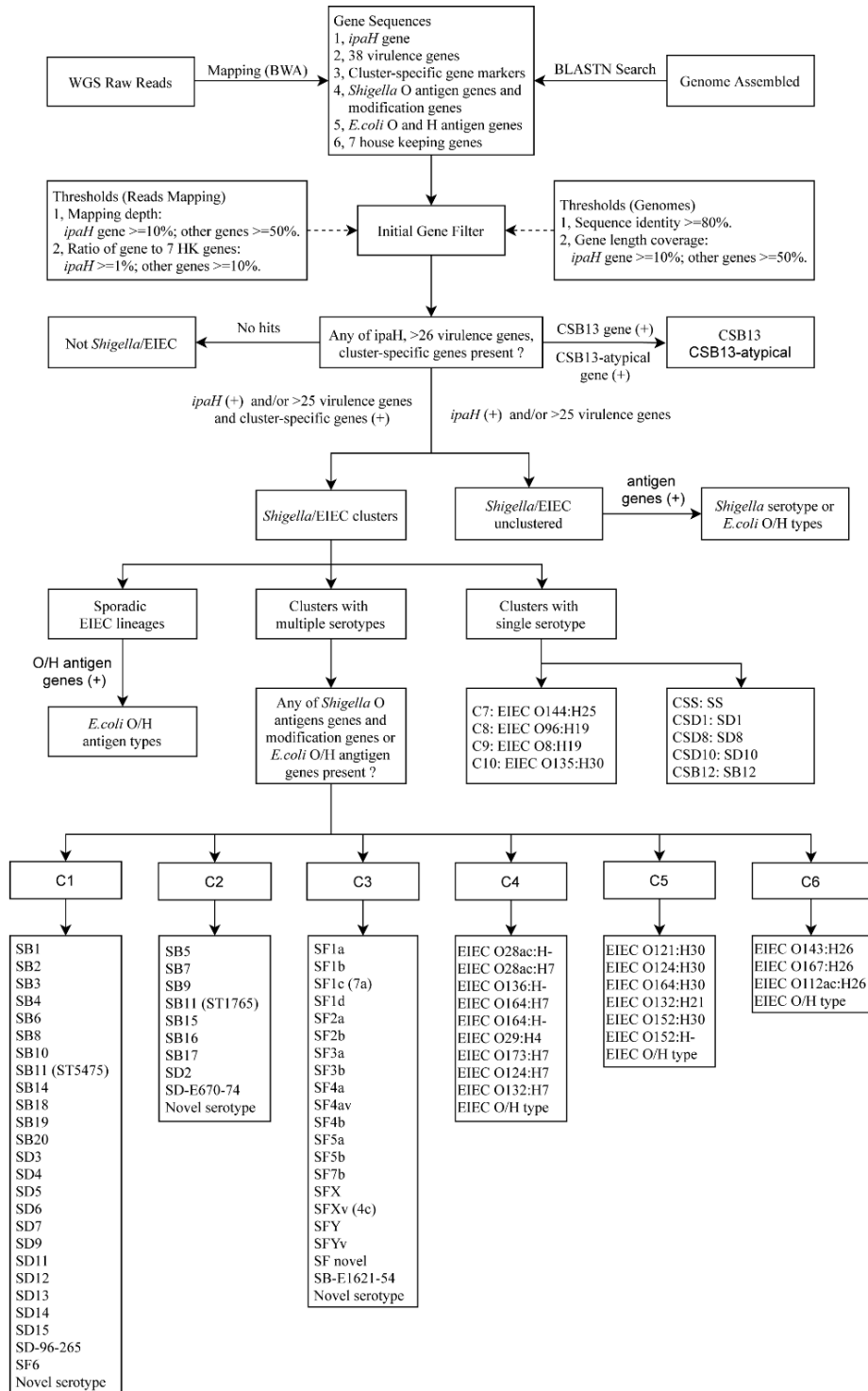
#### 4.5.5 Validation of cluster-specific gene markers

The ability of cluster-specific gene markers to correctly assign *Shigella* and EIEC isolates was evaluated with 15,501 *Shigella* and EIEC isolates in the validation dataset and 12,743 isolates from non-enteroinvasive *E. coli* control database. Using cluster-specific gene markers, 15,442 of the 15,501 (99.63%) *Shigella* and EIEC isolates were assigned to a single cluster which included 15,336 *Shigella* isolates, 102 EIEC isolates, 4 sporadic EIEC isolates. However, 38 (0.24%) isolates were assigned with more than one clusters and 21 isolates were not assigned to any of the identified clusters.

To confirm the cluster assignment by cluster-specific gene markers, we have divided the 15,501 validation isolates into 7 subgroups as it was impractical to construct a tree with all 15,501 genomes. We then constructed 7 “validation” phylogenetic trees (Fig. S4) using each of the 7 subgroups’ isolates and a subset of 575 isolates from the identification dataset consisting of 485 isolates representing each cluster, 72 ECOR isolates and 18 *E. albertii* strains. The cluster identity of an “validation” isolate was confirmed if the isolate was found within a branch that exclusively contained identification dataset isolates from that cluster and that branch had a bootstrap support value of 80% or greater (Fig. S4). The 7 phylogenies of 15,501 validation isolates showed that all 15,501 isolates were assigned to expected clusters with the exception of 4 isolates which were not grouped with any of the identified clusters (Table S2 column E).

Compared to cluster assignment by phylogenetic trees as the ground truth, cluster-specific gene markers assigned 15,442 of the 15,501 (99.63%) *Shigella* and EIEC isolates correctly to clusters and correctly identified 3 of the 21 isolates without cluster assignments. The accuracy of cluster assignments by cluster-specific gene markers was 99.64%. The sensitivity and specificity for each cluster-specific gene marker set for validation dataset were listed in Table S4.

We tested cluster-specific gene markers with the 12,743 non-enteroinvasive *E. coli* isolates. The *Shigella* and EIEC cluster-specific gene markers were highly specific with specificity varying from 98.8% to 100% for cluster-specific gene markers and 97.02% to 100% for sporadic EIEC specific gene markers. Details are listed in Table S4.



**Chapter 4. Figure 2: *in silico* serotyping pipeline workflow.** Schematic of *in silico* serotyping *Shigella* and EIEC by cluster-specific genes combined with the *ipaH* gene and O antigen and modification genes and H antigen genes, implemented in ShigEiFinder. Both assembled genomes and raw reads are accepted as data input.

#### 4.5.6 Development of an automated pipeline for molecular serotyping of *Shigella* and EIEC

Above results showed that cluster-specific gene markers were sensitive and specific and can distinguish *Shigella* and EIEC isolates. Therefore, we used these gene markers combined with established *Shigella* and EIEC serotype specific O and H antigen genes to develop an automated pipeline for *in silico* serotyping of *Shigella* and EIEC (Fig. 2). The pipeline is named *Shigella* EIEC Cluster Enhanced Serotype Finder (ShigEiFinder). ShigEiFinder can process either paired end Illumina sequencing reads or assembled genomes (installable package: <https://github.com/LanLab/ShigEiFinder>, online tool: <https://mgtdb.unsw.edu.au/ShigEiFinder/>). Details of the performance and algorithms incorporated into the ShigEiFinder are documented in the Data S2.

ShigEiFinder classifies isolates into Not *Shigella*/EIEC, *Shigella* or EIEC clusters, and *Shigella* or EIEC unclustered, based on the presence of *ipaH* gene, the number of virulence genes, cluster specific gene markers. The “Not *Shigella*/EIEC” assignment was determined by the absence of the *ipaH* gene, virulence genes (<26) and the absence of cluster-specific gene markers. The “*Shigella* or EIEC clusters” assignment was made based on the presence of *ipaH* gene, and/or more than 25 virulence genes together with the presence of any of cluster-specific gene markers or marker set, whereas the presence of *ipaH* gene and/or more than 25 virulence genes with absence of any of cluster-specific gene markers were assigned as “*Shigella* or EIEC unclustered”.

*Shigella* and EIEC isolates were differentiated and serotypes were assigned after cluster assignment. ShigEiFinder predicts a serotype through examining the presence of any of established *Shigella* serotype specific O antigen and modification genes and *E. coli* O and H antigen genes that differentiate the serotypes as ShigaTyper and SerotypeFinder (11, 63). A “novel serotype” is assigned if there is no match to known serotypes.

Two pairs of *Shigella* serotypes, SB1/SB20 and SB6/SB10, are known to be difficult to differentiate as they share identical O antigen genes (11, 46, 74). ShigaTyper used a heparinase gene for the differentiation of SB20 from SB1 and *wbaM* gene for the separation of SB6 from SB10. We found that fragments of the heparinase and *wbaM*

genes may be present in other serotypes and cannot accurately differentiate SB1/SB20 and SB6/SB10. We identified a SB20 specific gene which encoded a hypothetical protein with unknown function and located on a plasmid by comparative genomic analysis of all isolates in C1 accessory genome. The SB20 specific gene can reliably differentiate SB20 from SB1 and also one SNP each in *wzx* and *wzy* genes that can differentiate SB6 from SB10. We used these differences (Data S2) in ShigEiFinder for the prediction of these serotypes.

#### **4.5.7 The accuracy and specificity of ShigEiFinder in cluster typing**

The accuracy of ShigEiFinder was tested with 1,969 isolates (1,969 assembled genomes and 1,951 Illumina reads [note no reads available for 18 EIEC isolates from NCBI]) from the identification dataset and 15,501 isolates (15,501 assembled genomes and 15,501 Illumina reads) from the validation dataset. The results are listed in Table 3.

ShigEiFinder was able to assign 99.54% and 99.28% of the isolates in the identification dataset to clusters for assembled genomes and read mapping respectively. The accuracy was 99.70% and 99.81% for assembled genomes and read mapping respectively when applied to the validation dataset. Discrepancies were observed between assembled genomes and read mapping (Table 3). There were more isolates assigned to “*Shigella* or EIEC unclustered” in read mapping, in contrast there were more isolates assigned to multiple clusters in genome assemblies. The specificity of ShigEiFinder was 99.40% for assembled genomes and 99.38% for read mapping when evaluated with 12,743 non-*Shigella*/EIEC *E. coli* isolates. An additional 2 isolates were detected as sporadic EIEC lineages by read mapping.

**Chapter 4. Table 3: The accuracy of ShigEiFinder with identification dataset and validation dataset**

ShigEiFinder assignments	Identification Dataset (n=1,969) <sup>a</sup>		Validation dataset (n=15,501)	
	Genomes	Reads mapping	Genomes	Reads mapping
<i>Shigella</i> or EIEC clusters	1871	1848	15,455	15,471
Multiple <i>Shigella</i> or EIEC clusters	9	6	33	7
<i>Shigella</i> or EIEC unclustered	0	8	13	23
Not <i>Shigella</i> /EIEC	89	89	0	0
Accuracy <sup>b</sup>	99.54%	99.28%	99.70%	99.81%

<sup>a</sup>: Reads were not available for 18 EIEC isolates downloaded from NCBI in identification dataset. Identification dataset has 90 non-*Shigella*/EIEC isolates including 72 ECOR isolates and 18 *E.albertii* isolates. One of *E.albertii* isolate was assigned to SB13 by ShigaTyper which was grouped into SB13 cluster on the phylogenetic tree. <sup>b</sup>: The accuracy was defined as the number of *Shigella* and EIEC isolates being correctly assigned to cluster over the total number of tested.

Chapter 4. Table 4: The assignments of 15,501 validation isolates by ShigEiFinder and Shigatyper

ShigEiFinder assignment	ShigaTyper assignment				Total
	Agreement with ShigEiFinder	Discrepant with ShigEiFinder			
		<i>Shigella</i>	EIEC	Non-assignment*	
SS	1,515	0	7,465	19	8,999
SF	4,644	0	117	71	4,832
C1 and C2 (SB and SD)	1,004	0	17	151	1,172
SB12	4	0	0	2	6
SB13	1	0	0	0	1
SB13-atypical	2	0	0	0	2
SD1	80	0	244	2	326
SD8	2	0	1	0	3
SD10	0	0	0	1	1
EIEC	101	1	0	0	102
Sporadic EIEC lineages	0	1	15	0	16
Multiple clusters	0	0	5	2	7
<i>Shigella</i> or EIEC unclustered	0	23	11	0	34
Total	7,353	25	7,875	248	15,501

\*: Non-assignment: multiple *wzx* genes and non-prediction.

#### 4.5.8 Comparison of ShigEiFinder and ShigaTyper

To demonstrate ShigEiFinder for differentiation of *Shigella* from EIEC and enhancement of cluster based serotyping, the comparison of read mapping results between ShigEiFinder and the existing *in silico* *Shigella* identification pipeline ShigaTyper (11) was performed. Since ShigaTyper recommends the use of read mapping, we compared ShigEiFinder read mapping results with ShigaTyper read mapping results.

The 488 isolates used in Wu et al (11) were tested using ShigEiFinder. These 488 isolates consisted of 25 EIEC isolates, 420 *Shigella* isolates and 45 non-*Shigella*/EIEC isolates. The assignment of 477 of 488 isolates by ShigEiFinder was in agreement with that by ShigaTyper. Of the remaining 11 isolates (1 EIEC isolate and 10 *Shigella* isolates), 2 *Shigella* isolates were assigned to EIEC and 8 *Shigella* isolates and 1 EIEC isolate were untypeable (either multiple *wzx* or no *wzx* genes found) by ShigaTyper, whereas 1 EIEC isolate was assigned to EIEC (C4) and 10 *Shigella* isolates were assigned to *Shigella* clusters by ShigEiFinder.

The read mapping results for 15,501 *Shigella* and EIEC isolates from validation dataset were then compared. ShigEiFinder assigned 15,460 of 15,501 *Shigella* and EIEC isolates to *Shigella* or EIEC clusters and then to a serotype. By contrast, ShigaTyper assigned 7,277 isolates to *Shigella*, 7,976 isolates to EIEC, 177 isolates to multiple *wzx* genes and failed to type 71 isolates. The total of 7,353 isolates predicted as *Shigella* (7,252) or EIEC (101) by ShigaTyper agreed with the results of ShigEiFinder (Table 4). For the 8,148 isolates typed as EIEC or untypeable by ShigaTyper, 8,107 isolates were assigned to *Shigella* or EIEC clusters by ShigEiFinder (Table 4). Of these isolates, the majority belonged to SS, SD1 and SF which were erroneously predicted as EIEC by ShigaTyper.

Compared to the phylogenetic analysis results of cluster identity of the isolates as ground truth, ShigEiFinder have 99.74% (15,460/15,501) accuracy to differentiate *Shigella* isolates from EIEC. While ShigaTyper assigned only 47.6% isolates correctly in the same dataset we tested.



## 4.6 Discussion

### 4.6.1 Determining phylogenetic clusters for better separation of *Shigella* isolates from EIEC

From a phylogenetic perspective, *Shigella* and EIEC strains consisted of multiple phylogenetic lineages derived from commensal *E. coli*, which do not reflect the taxonomic classification of *Shigella* as a genus (23, 25, 26, 28, 38, 41). In the present study, we identified all phylogenetic clusters of *Shigella* and EIEC through large scale examination of publicly available genomes. Phylogenetic results demonstrated that *Shigella* isolates had at least 10 clusters while EIEC isolates had at least 7 clusters. The 10 *Shigella* clusters included the 8 previously defined lineages including 3 major clusters (C1, C2 and C3) and 5 outliers (SD1, SD8, SD10, SB13 and SS) (25) and 2 newly identified clusters (SB12 and SB13-atypical). The 7 EIEC clusters consisted of 4 previously defined EIEC clusters (C4, C5, C6 and C7) (26) and 3 newly identified EIEC clusters (C8, C9 and C10).

Our WGS-based phylogeny provided high resolution for assigning *Shigella* and EIEC isolates to clusters. Several serotypes that are currently increasing in frequency (SB19, SB20, SD14, SD15, SD provisional serotype 96-626) (75-78) were assigned to clusters and five new clusters/outliers were identified. Newly identified clusters C8 (EIEC O96:H19) and C9 (EIEC O8:H19) represented the emergence of novel EIEC serotypes. A recent study revealed that EIEC serotype O96:H19 (C8) could be the result of a recent acquisition of the invasion plasmid by commensal *E. coli* (79). The EIEC serotype O8:H19 (C9) had not been reported previously.

Apart from the 17 major clusters of *Shigella* and EIEC, the presence of 53 sporadic EIEC lineages indicated greater genetic diversity than has been observed previously. Isolates belonging to these sporadic EIEC lineages were more closely related to non-enteroinvasive *E. coli* isolates than to major *Shigella* and EIEC lineages. However, 41 of these isolates, representing 38 sporadic EIEC lineages, carried pINV. *Shigella* and EIEC both carry the *Shigella* virulence plasmid pINV which is vital for virulence and distinguishes *Shigella* and EIEC from other *E. coli* (24, 32, 68). Therefore, these isolates may represent recently formed EIEC lineages through acquisition of the pINV. The remaining 18 isolates contained the *ipaH* gene but may or may not carry pINV. It is

possible that these strains carried very low copy number of the pINV or the pINV plasmid was lost during isolation or culture.

#### **4.6.2 Highly sensitive and specific cluster-specific gene markers for differentiation of *Shigella* and EIEC isolates**

The cluster-specific gene marker sets can be used to differentiate *Shigella* and EIEC from non-enteroinvasive *E. coli* independent of the presence of *ipaH* gene. The *ipaH* gene as a molecular target has been used to differentiate *Shigella* and EIEC from non-enteroinvasive *E. coli* (24, 43-45). In our study, the cluster-specific gene markers were specific to *Shigella* and EIEC with 98.8% to 100% specificity when evaluated on non-enteroinvasive *E. coli* control database, providing confidence that the cluster-specific genes or sets are robust markers for the identification of *Shigella* and EIEC.

Several studies have identified phylogenetic related genomic markers for discrimination of *Shigella* and EIEC (23, 27, 28, 41, 55, 56). However, these phylogenetic analyses were performed only with a small number of genomes (23, 28, 55). In addition, non-invasive *E. coli* isolates were included in some of the phylogenetic clusters identified (28) which led to non-invasive *E. coli* isolates being identified by the markers. We identified cluster-specific gene markers for each respective cluster which were exclusively composed of *Shigella* or EIEC isolates. A previous study identified 6 loci to distinguish EIEC from *Shigella* (23). We searched the 6 loci against our *Shigella* and EIEC database and found that some *Shigella* isolates were misidentified as EIEC, such as SD8 isolates were incorrectly identified as EIEC subtype 13. Our cluster-specific genes can differentiate SD8 from EIEC with 100% accuracy. Overall, the cluster-specific gene marker sets described here provided nearly perfect differentiation of *Shigella* from EIEC.

The cluster-specific gene marker sets can differentiate SS and SF (with exception of SF6) from SB and SD. SF and SS are the major cause of *Shigella* infections, accounting for up to 89.6% annual cases (10, 12, 13). Differentiation of SS and SF isolates from SB and SD is also beneficial for diagnosis and surveillance. A recent study identified “species” specific markers for the detection of each of the four *Shigella* “species” and validated with only one isolate per species (55). Whereas a set of SF specific genes and SS specific

genes in our study can correctly identify SF isolate and SS isolates with 99.64% accuracy when applied to 15,501 *Shigella* and EIEC isolates.

It should be noted that we were unable to validate cluster-specific gene markers of C6, C7, C10 and CSD10. These clusters are rare and once isolates were included in the identification dataset, none remained for validation. Therefore, these markers for the C6, C7, C10 and SD10 clusters are tentative and require future validation when more genomes are available. Genes specific to each of the 53 sporadic EIEC lineages were also based on very small number of genomes and should be used with caution. However, since these sporadic lineages are very low in frequency, they may be rarely encountered in practice and thus have relatively little effect on the overall applicability of the lineage specific markers to *Shigella* and EIEC typing.

#### **4.6.3 ShigEiFinder can accurately type *Shigella* and EIEC**

ShigEiFinder can accurately differentiate *Shigella* from EIEC whereas there were a large proportion of isolates incorrectly assigned by ShigaTyper. The majority of the isolates predicted as EIEC by ShigaTyper were SS or SD1 as they belonged to SS and SD1 specific STs and were positive to a set of SS or SD1 specific gene markers and grouped into SS or SD1 cluster on our phylogenetic tree. The genes used in ShigaTyper were SS specific marker Ss\_methylase gene (80, 81) together with SS O antigen wzx gene. However, SS specific marker Ss\_methylase gene was found in other *Shigella* serotypes and EIEC (11) and SS O antigen wzx gene were located on a plasmid which is frequently lost (82). Similarly, the SD1 O antigen genes used in ShigaTyper were plasmid-borne which may also lead to inconsistent detection (83, 84). By contrast, the cluster-specific gene markers used in ShigEiFinder for identification of *Shigella* and EIEC provided higher discriminatory power than ShigaTyper.

ShigEiFinder was able to serotype over 59 *Shigella* serotypes and 22 EIEC serotypes. ShigEiFinder can assign *Shigella* and EIEC isolates to serotype level using cluster specific markers to enhance the accuracy. For clusters containing more than one serotype including the major *Shigella* and EIEC clusters C1-C6, once an isolate is assigned to a cluster, only serotype associated O antigen and modification genes found in that cluster need be examined. This allows the elimination of ambiguous or incorrect serotype

assignments that may otherwise occur, increasing the overall accuracy of the method. For the cluster contain only one serotype such as SD1, SD8, SD10, SB13, SB12, EIEC C7-C10, cluster specific markers can also be used a proxy to serotyping but with increased robustness when the combination of cluster-specific gene marker combined with serotype associated O antigen and modification genes was used.

ShigEiFinder will be useful for clinical, epidemiological and diagnostic investigations and the cluster-specific gene markers identified could be adapted for metagenomics or culture independent typing.

## **4.7 Conclusion**

This study analysed over 17,000 publicly available *Shigella* and EIEC isolates and identified 10 clusters of *Shigella*, 7 clusters of EIEC and 53 sporadic types of EIEC. Cluster-specific gene marker sets for the 17 major clusters and 53 sporadic types were identified and found to be valuable for *in silico* typing. We additionally developed ShigEiFinder, a freely available *in silico* serotyping pipeline incorporating the cluster-specific gene markers to facilitate serotyping of *Shigella* and EIEC isolates using genome sequences with very high specificity and sensitivity.

## **4.8 Authors and contributors**

Conceptualization: R.L, M.P.; Investigation: X.Z., M.P., T.N., S.K.; Methodology: M.P., R.L., X.Z; Writing – original draft: X.Z.; Writing – review and editing: M.P., R.L.

## **4.9 Acknowledgements**

The authors thank Duncan Smith and Robin Heron from UNSW Research Technology Services for computing assistance.

## **4.10 Data bibliography**

Zhang X, Payne M, Nguyen T, Kaur S, Lan R. All the sequencing data generated within this study, NCBI BioProject number (PRJNA692536).

## 4.11 Abbreviations

SS, *Shigella sonnei*; SF, *Shigella flexneri*; SB, *Shigella boydii*; SD, *Shigella dysenteriae*; EIEC, Enteroinvasive *Escherichia coli*; NCBI SRA, National Center for Biotechnology Information Sequence Read Archive; ST, sequence type; rST, ribosomal ST; MLST, Multilocus sequence typing; rMLST, Ribosomal MLST; ECOR, *Escherichia coli* reference collection; WGS, whole-genome sequencing; TP, true positive; FN, false negative; FP, false positive; HK, House Keeping.

## 4.12 References

1. DuPont HL, Levine MM, Hornick RB, Formal SB. Inoculum size in shigellosis and implications for expected mode of transmission. *The Journal of infectious diseases*. 1989;159(6):1126-8.
2. Troeger C, Forouzanfar M, Rao PC, Khalil I, Brown A, Reiner Jr RC, et al. Estimates of global, regional, and national morbidity, mortality, and aetiologies of diarrhoeal diseases: a systematic analysis for the Global Burden of Disease Study 2015. *The Lancet Infectious Diseases*. 2017;17(9):909-48.
3. Kirk MD, Pires SM, Black RE, Caipo M, Crump JA, Devleesschauwer B, et al. World Health Organization Estimates of the Global and Regional Disease Burden of 22 Foodborne Bacterial, Protozoal, and Viral Diseases, 2010: A Data Synthesis. *PLoS medicine*. 2015;12(12):e1001921.
4. World HO. Guidelines for the control of shigellosis, including epidemics due to *Shigella dysenteriae* type 1. 2005.
5. Brengi SP, Sun Q, Bolaños H, Duarte F, Jenkins C, Pichel M, et al. PCR-Based Method for *Shigella flexneri* Serotyping: International Multicenter Validation. *J Clin Microbiol*. 2019;57(4).
6. Khalil IA, Troeger C, Blacker BF, Rao PC, Brown A, Atherly DE, et al. Morbidity and mortality due to *shigella* and enterotoxigenic *Escherichia coli* diarrhoea: the Global Burden of Disease Study 1990-2016. *The Lancet Infectious diseases*. 2018;18(11):1229-40.
7. Kotloff KL, Nataro JP, Blackwelder WC, Nasrin D, Farag TH, Panchalingam S, et al. Burden and aetiology of diarrhoeal disease in infants and young children in

developing countries (the Global Enteric Multicenter Study, GEMS): a prospective, case-control study. *Lancet*. 2013;382(9888):209-22.

8. van den Beld MJC, Warmelink E, Friedrich AW, Reubsat FAG, Schipper M, de Boer RF, et al. Incidence, clinical implications and impact on public health of infections with *Shigella spp.* and entero-invasive *Escherichia coli* (EIEC): results of a multicenter cross-sectional study in the Netherlands during 2016-2017. *BMC Infect Dis*. 2019;19(1):1037.

9. Edwards PR, Ewing WH. Identification of enterobacteriaceae. *Identification of Enterobacteriaceae*. 1972(Third edition).

10. The HC, Thanh DP, Holt KE, Thomson NR, Baker S. The genomic signatures of *Shigella* evolution, adaptation and geographical spread. *Nature reviews Microbiology*. 2016;14(4):235-50.

11. Wu Y, Lau HK, Lee T, Lau DK, Payne J. *In Silico* Serotyping Based on Whole-Genome Sequencing Improves the Accuracy of *Shigella* Identification. *Applied and environmental microbiology*. 2019;85(7).

12. Livio S, Strockbine NA, Panchalingam S, Tennant SM, Barry EM, Marohn ME, et al. *Shigella* isolates from the global enteric multicenter study inform vaccine development. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America*. 2014;59(7):933-41.

13. Group OW. Monitoring the incidence and causes of diseases potentially transmitted by food in Australia: Annual report of the OzFoodNet network, 2011. *Communicable diseases intelligence quarterly report*. 2015;39(2):E236.

14. Connor TR, Barker CR, Baker KS, Weill FX, Talukder KA, Smith AM, et al. Species-wide whole genome sequencing reveals historical global spread and recent local persistence in *Shigella flexneri*. *Elife*. 2015;4:e07335.

15. Gomes TA, Elias WP, Scaletsky IC, Guth BE, Rodrigues JF, Piazza RM, et al. Diarrheagenic *Escherichia coli*. *Braz J Microbiol*. 2016;47 Suppl 1(Suppl 1):3-30.

16. Levine MM. *Escherichia coli* that cause diarrhea: enterotoxigenic, enteropathogenic, enteroinvasive, enterohemorrhagic, and enteroadherent. *The Journal of infectious diseases*. 1987;155(3):377-89.

17. Tai AY, Easton M, Encena J, Rotty J, Valcanis M, Howden BP, et al. A review of the public health management of shigellosis in Australia in the era of culture-independent

- diagnostic testing. *Australian and New Zealand journal of public health*. 2016;40(6):588-91.
18. Taylor D, Echeverria P, Sethabutr O, Pitarangsi C, Leksomboon U, Blacklow N, et al. Clinical and microbiologic features of *Shigella* and enteroinvasive *Escherichia coli* infections detected by DNA hybridization. *Journal of clinical microbiology*. 1988;26(7):1362-6.
  19. Pasqua M, Michelacci V, Di Martino ML, Tozzoli R, Grossi M, Colonna B, et al. The Intriguing Evolutionary Journey of Enteroinvasive *E. coli* (EIEC) toward Pathogenicity. *Frontiers in microbiology*. 2017;8:2390.
  20. Herzig CTA, Fleischauer AT, Lackey B, Lee N, Lawson T, Moore ZS, et al. Notes from the Field: Enteroinvasive *Escherichia coli* Outbreak Associated with a Potluck Party - North Carolina, June-July 2018. *MMWR Morbidity and mortality weekly report*. 2019;68(7):183-4.
  21. Pettengill EA, Hoffmann M, Binet R, Roberts RJ, Payne J, Allard M, et al. Complete Genome Sequence of Enteroinvasive *Escherichia coli* O96:H19 Associated with a Severe Foodborne Outbreak. *Genome announcements*. 2015;3(4).
  22. Escher M, Scavia G, Morabito S, Tozzoli R, Maugliani A, Cantoni S, et al. A severe foodborne outbreak of diarrhoea linked to a canteen in Italy caused by enteroinvasive *Escherichia coli*, an uncommon agent. *Epidemiology and infection*. 2014;142(12):2559-66.
  23. Dhakal R, Wang Q, Lan R, Howard P, Sintchenko V. Novel multiplex PCR assay for identification and subtyping of enteroinvasive *Escherichia coli* and differentiation from *Shigella* based on target genes selected by comparative genomics. *J Med Microbiol*. 2018;67(9):1257-64.
  24. van den Beld MJ, Reubsat FA. Differentiation between *Shigella*, enteroinvasive *Escherichia coli* (EIEC) and noninvasive *Escherichia coli*. *European journal of clinical microbiology & infectious diseases : official publication of the European Society of Clinical Microbiology*. 2012;31(6):899-904.
  25. Pupo GM, Lan R, Reeves PR. Multiple independent origins of *Shigella* clones of *Escherichia coli* and convergent evolution of many of their characteristics. *Proc Natl Acad Sci U S A*. 2000;97(19):10567-72.

26. Lan R, Alles MC, Donohoe K, Martinez MB, Reeves PR. Molecular evolutionary relationships of enteroinvasive *Escherichia coli* and *Shigella* spp. *Infection and immunity*. 2004;72(9):5080-8.
27. Sahl JW, Morris CR, Emberger J, Fraser CM, Ochieng JB, Juma J, et al. Defining the phylogenomics of *Shigella* species: a pathway to diagnostics. *Journal of clinical microbiology*. 2015;53(3):951-60.
28. Pettengill EA, Pettengill JB, Binet R. Phylogenetic Analyses of *Shigella* and Enteroinvasive *Escherichia coli* for the Identification of Molecular Epidemiological Markers: Whole-Genome Comparative Analysis Does Not Support Distinct Genera Designation. *Frontiers in microbiology*. 2015;6:1573.
29. Cheasty T, Rowe B. Antigenic relationships between the enteroinvasive *Escherichia coli* O antigens O28ac, O112ac, O124, O136, O143, O144, O152, and O164 and *Shigella* O antigens. *Journal of clinical microbiology*. 1983;17(4):681-4.
30. Landersjö C, Weintraub A, Widmalm G. Structure determination of the O-antigen polysaccharide from the enteroinvasive *Escherichia coli* (EIEC) O143 by component analysis and NMR spectroscopy. *Carbohydr Res*. 1996;291:209-16.
31. Linnerborg M, Weintraub A, Widmalm G. Structural studies of the O-antigen polysaccharide from the enteroinvasive *Escherichia coli* O164 cross-reacting with *Shigella dysenteriae* type 3. *Eur J Biochem*. 1999;266(2):460-6.
32. Lan R, Lumb B, Ryan D, Reeves PR. Molecular evolution of large virulence plasmid in *Shigella* clones and enteroinvasive *Escherichia coli*. *Infection and immunity*. 2001;69(10):6303-9.
33. Sansonetti PJ, d'Hauteville H, Ecobichon C, Pourcel C. Molecular comparison of virulence plasmids in *Shigella* and enteroinvasive *Escherichia coli*. *Annales de microbiologie*. 1983;134a(3):295-318.
34. Hale TL. Genetic basis of virulence in *Shigella* species. *Microbiological reviews*. 1991;55(2):206-24.
35. Venkatesan MM, Buysse JM, Kopecko DJ. Use of *Shigella flexneri* *ipaC* and *ipaH* gene sequences for the general identification of *Shigella* spp. and enteroinvasive *Escherichia coli*. *J Clin Microbiol*. 1989;27(12):2687-91.
36. Jin Q, Yuan Z, Xu J, Wang Y, Shen Y, Lu W, et al. Genome sequence of *Shigella flexneri* 2a: insights into pathogenicity through comparison with genomes of *Escherichia coli* K12 and O157. *Nucleic Acids Res*. 2002;30(20):4432-41.



37. Yang F, Yang J, Zhang X, Chen L, Jiang Y, Yan Y, et al. Genome dynamics and diversity of *Shigella* species, the etiologic agents of bacillary dysentery. *Nucleic acids research*. 2005;33(19):6445-58.
38. Yang J, Nie H, Chen L, Zhang X, Yang F, Xu X, et al. Revisiting the molecular evolutionary history of *Shigella* spp. *Journal of molecular evolution*. 2007;64(1):71-9.
39. Hyma KE, Lacher DW, Nelson AM, Bumbaugh AC, Janda JM, Strockbine NA, et al. Evolutionary genetics of a new pathogenic *Escherichia* species: *Escherichia albertii* and related *Shigella boydii* strains. *Journal of bacteriology*. 2005;187(2):619-28.
40. Walters LL, Raterman EL, Grys TE, Welch RA. Atypical *Shigella boydii* 13 encodes virulence factors seen in attaching and effacing *Escherichia coli*. *FEMS microbiology letters*. 2012;328(1):20-5.
41. Hazen TH, Leonard SR, Lampel KA, Lacher DW, Maurelli AT, Rasko DA. Investigating the Relatedness of Enteroinvasive *Escherichia coli* to Other *E. coli* and *Shigella* Isolates by Using Comparative Genomics. *Infection and immunity*. 2016;84(8):2362-71.
42. Silva RM, Toledo MR, Trabulsi LR. Biochemical and cultural characteristics of invasive *Escherichia coli*. *Journal of clinical microbiology*. 1980;11(5):441-4.
43. de Boer RF, Ott A, Kesztyüs B, Kooistra-Smid AM. Improved detection of five major gastrointestinal pathogens by use of a molecular screening approach. *Journal of clinical microbiology*. 2010;48(11):4140-6.
44. van den Beld MJC, Friedrich AW, van Zanten E, Reubsaet FAG, Kooistra-Smid M, Rossen JWA. Multicenter evaluation of molecular and culture-dependent diagnostics for *Shigella* species and Entero-invasive *Escherichia coli* in the Netherlands. *Journal of microbiological methods*. 2016;131:10-5.
45. Van Lint P, De Witte E, Ursi J, Van Herendael B, Van Schaeren J. A screening algorithm for diagnosing bacterial gastroenteritis by real-time PCR in combination with guided culture. *Diagnostic microbiology*. 2016;85(2):255-9.
46. Liu B, Knirel YA, Feng L, Perepelov AV, Senchenkova SN, Wang Q, et al. Structure and genetics of *Shigella* O antigens. *FEMS microbiology reviews*. 2008;32(4):627-53.
47. Cai H, Lu L, Muckle C, Prescott J, Chen S. Development of a novel protein microarray method for serotyping *Salmonella enterica* strains. *Journal of clinical microbiology*. 2005;43(7):3427-30.

48. Wattiau P, Boland C, Bertrand S. Methodologies for *Salmonella enterica* subsp. *enterica* subtyping: gold standards and alternatives. *Applied and environmental microbiology*. 2011;77(22):7877-85.
49. Li Y, Cao B, Liu B, Liu D, Gao Q, Peng X, et al. Molecular detection of all 34 distinct O-antigen forms of *Shigella*. *J Med Microbiol*. 2009;58(Pt 1):69-81.
50. Sun Q, Lan R, Wang Y, Zhao A, Zhang S, Wang J, et al. Development of a multiplex PCR assay targeting O-antigen modification genes for molecular serotyping of *Shigella flexneri*. *Journal of clinical microbiology*. 2011;49(11):3766-70.
51. van der Ploeg CA, Rogé AD, Bordagorria XL, de Urquiza MT, Castillo ABC, Bruno SB. Design of Two Multiplex PCR Assays for Serotyping *Shigella flexneri*. *Foodborne pathogens and disease*. 2018;15(1):33-8.
52. van den Beld MJC, de Boer RF, Reubsaet FAG, Rossen JWA, Zhou K, Kuiling S, et al. Evaluation of a Culture-Dependent Algorithm and a Molecular Algorithm for Identification of *Shigella spp.*, *Escherichia coli*, and Enteroinvasive *E. coli*. *J Clin Microbiol*. 2018;56(10).
53. Løbersli I, Wester AL, Kristiansen Å, Brandal LT. Molecular Differentiation of *Shigella Spp.* from Enteroinvasive *E. Coli*. *European journal of microbiology & immunology*. 2016;6(3):197-205.
54. Pavlovic M, Luze A, Konrad R, Berger A, Sing A, Busch U, et al. Development of a duplex real-time PCR for differentiation between *E. coli* and *Shigella spp.* *Journal of applied microbiology*. 2011;110(5):1245-51.
55. Kim HJ, Ryu JO, Song JY, Kim HY. Multiplex Polymerase Chain Reaction for Identification of Shigellae and Four *Shigella* Species Using Novel Genetic Markers Screened by Comparative Genomics. *Foodborne pathogens and disease*. 2017;14(7):400-6.
56. Chattaway MA, Schaefer U, Tewolde R, Dallman TJ, Jenkins C. Identification of *Escherichia coli* and *Shigella* Species from Whole-Genome Sequences. *Journal of clinical microbiology*. 2017;55(2):616-23.
57. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods*. 2017;14(4):417-9.
58. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome biology*. 2014;15(3):R46.

59. Alikhan NF, Zhou Z, Sergeant MJ, Achtman M. A genomic overview of the population structure of *Salmonella*. *PLoS Genet*. 2018;14(4):e1007261.
60. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of computational biology : a journal of computational molecular cell biology*. 2012;19(5):455-77.
61. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics*. 2013;29(8):1072-5.
62. Jolley KA, Maiden MC. BIGSdb: Scalable analysis of bacterial genome variation at the population level. *BMC Bioinformatics*. 2010;11:595.
63. Joensen KG, Tetzschner AM, Iguchi A, Aarestrup FM, Scheutz F. Rapid and Easy *In Silico* Serotyping of *Escherichia coli* Isolates by Use of Whole-Genome Sequencing Data. *Journal of clinical microbiology*. 2015;53(8):2410-26.
64. Hu D, Liu B, Wang L, Reeves PR. Living Trees: High-Quality Reproducible and Reusable Construction of Bacterial Phylogenetic Trees. *Mol Biol Evol*. 2020;37(2):563-75.
65. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res*. 2019;47(W1):W256-w9.
66. Zhou Z, Alikhan NF, Sergeant MJ, Luhmann N, Vaz C, Francisco AP, et al. GrapeTree: visualization of core genomic relationships among 100,000 bacterial pathogens. *Genome Res*. 2018;28(9):1395-404.
67. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv*. 2013.
68. Buchrieser C, Glaser P, Rusniok C, Nedjari H, D'Hauteville H, Kunst F, et al. The virulence plasmid pWR100 and the repertoire of proteins secreted by the type III secretion apparatus of *Shigella flexneri*. *Molecular microbiology*. 2000;38(4):760-71.
69. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009;25(16):2078-9.
70. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. 2010;26(6):841-2.
71. Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics*. 2014;30(14):2068-9.

72. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MT, et al. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics*. 2015;31(22):3691-3.
73. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. *BMC Bioinformatics*. 2009;10:421.
74. Senchenkova SN, Feng L, Yang J, Shashkov AS, Cheng J, Liu D, et al. Structural and genetic characterization of the *Shigella boydii* type 10 and type 6 O antigens. *Journal of bacteriology*. 2005;187(7):2551-4.
75. Ansaruzzaman M, Kibriya A, Rahman A, Neogi P, Faruque A, Rowe B, et al. Detection of provisional serovars of *Shigella dysenteriae* and designation as *S. dysenteriae* serotypes 14 and 15. *Journal of clinical microbiology*. 1995;33(5):1423-5.
76. Balows A. Manual of clinical microbiology 8th edition: PR Murray, EJ Baron, JH Jorgenson, MA Tenover, and RH Tenover, eds., ASM Press, 2003, 2113 pages, 2 vol, 2003+ subject & author indices, ISBN: 1-555810255-4, US \$189.95. *Diagnostic Microbiology*. 2003;47(4):625.
77. Woodward DL, Clark CG, Caldeira RA, Ahmed R, Soule G, Bryden L, et al. Identification and characterization of *Shigella boydii* 20 serovar nov., a new and emerging *Shigella* serotype. *J Med Microbiol*. 2005;54(Pt 8):741-8.
78. Kim J, Lindsey RL, Garcia-Toledo L, Loparev VN, Rowe LA, Batra D, et al. High-Quality Whole-Genome Sequences for 59 Historical *Shigella* Strains Generated with PacBio Sequencing. *Genome Announc*. 2018;6(15).
79. Michelacci V, Prosseda G, Maugliani A, Tozzoli R, Sanchez S, Herrera-León S, et al. Characterization of an emergent clone of enteroinvasive *Escherichia coli* circulating in Europe. *Clinical Microbiology*. 2016;22(3):287. e11-. e19.
80. Liu J, Platts-Mills JA, Juma J, Kabir F, Nkeze J, Okoi C, et al. Use of quantitative molecular diagnostic methods to identify causes of diarrhoea in children: a reanalysis of the GEMS case-control study. *Lancet*. 2016;388(10051):1291-301.
81. Cho MS, Ahn TY, Joh K, Kwon OS, Jheong WH, Park DS. A novel marker for the species-specific detection and quantitation of *Shigella sonnei* by targeting a methylase gene. *J Microbiol Biotechnol*. 2012;22(8):1113-7.
82. Sansonetti PJ, Kopecko DJ, Formal SB. *Shigella sonnei* plasmids: evidence that a large plasmid is necessary for virulence. *Infection and immunity*. 1981;34(1):75-83.
83. Feng L, Perepelov AV, Zhao G, Shevelev SD, Wang Q, Senchenkova SN, et al. Structural and genetic evidence that the *Escherichia coli* O148 O antigen is the precursor

of the *Shigella dysenteriae* type 1 O antigen and identification of a glucosyltransferase gene. *Microbiology* (Reading, England). 2007;153(Pt 1):139-47.

84. Göhmann S, Manning P, Alpert C-A, Walker M, Timmis K. Lipopolysaccharide O-antigen biosynthesis in *Shigella dysenteriae* serotype 1: analysis of the plasmid-carried *rfp* determinant. *Microbial pathogenesis*. 1994;16(1):53-64.

# **Chapter 5. Improved genomic identification, clustering and serotyping of Shiga toxin-producing *Escherichia coli* using cluster/serotype-specific gene markers**

## **5.1 Link to thesis**

STEC infections have a significant impact on public health worldwide and detection and differentiation of STEC is vital for public health. I presented the limitations of current identification and serotyping methods for STEC in Chapter 1. I also took advantage of the large number of the genome sequences for STEC in public databases to perform genomic analysis for identification of robust genomic markers for accurate prediction and identification of STEC. This chapter addresses the fourth aim of this thesis.

I have submitted this work to Frontiers in Cellular and Infection Microbiology 08/09/2021: Zhang X, Payne M, Kaur S, Lan R. Improved genomic identification, clustering and serotyping of Shiga toxin-producing *Escherichia coli* using cluster/serotype-specific gene markers.

**The Supplementary Material for this article can be found online at:**

<https://drive.google.com/drive/folders/1HXJvKIHHYeQ9-ZoCQ7WnY7I3oX4lhUfW?usp=sharing>

**Supplemental material for this article is also listed at Appendix IV.**

## **5.2 Abstract**

Shiga toxin-producing *Escherichia coli* (STEC) have more than 470 serotypes. The well-known STEC O157:H7 serotype is a leading cause of STEC infections in humans. However, the incidence of non-O157:H7 STEC serotypes associated with foodborne outbreaks and human infections has increased in recent years. Current detection and serotyping assays are focusing on STEC O157:H7 and top 6 (“Big 6”) non-O157:H7

STEC serotypes. In this study, we performed phylogenetic analysis of nearly 41,000 publicly available STEC genomes representing 460 different STEC serotypes and identified 19 major and 229 minor STEC clusters. STEC cluster-specific gene markers were then identified through comparative genomic analysis. We further identified serotype-specific gene markers for the top 10 most frequent non-O157:H7 STEC serotypes. The gene markers had 99.54% accuracy and more than 97.25% specificity when tested using 38,534 STEC and 14,216 non-STEC *E. coli* genomes, respectively. Using shotgun metagenomic sequencing reads of STEC spiked food samples from a published study, we demonstrated that these gene markers can detect the spiked STEC serotype accurately. In addition, we developed a freely available *in silico* serotyping pipeline named STECFinder that combined these robust gene markers with established *E. coli* serotype specific O antigen genes and H antigen genes and *stx* genes for accurate identification, cluster determination and serotyping of STEC. STECFinder can assign 99.85% and 99.83% of 38,534 STEC isolates to STEC clusters using assembled genomes and Illumina reads respectively and simultaneously predict *stx* subtypes and STEC serotypes. The cluster/serotype-specific gene markers could be adapted for metagenomics based diagnosis and culture independent typing, facilitating rapid STEC identification. STECFinder is available as an installable package (<https://github.com/LanLab/STECFinder>) and will be useful for *in silico* STEC identification and typing using genome data.

**Running title:** *in silico* typing pipeline STECFinder

**Keywords:** STEC O157:H7, Non-O157:H7 STEC serotypes, STEC phylogenetic clusters, cluster/serotype-specific gene markers, STEC serotyping, *in silico* typing pipeline STECFinder, metagenomics

## 5.3 Introduction

Shiga toxin-producing *Escherichia coli* (STEC) are an important cause of foodborne disease worldwide (Tuttle et al., 1999; Teunis et al., 2008; World Health Organization, 2019). STEC causes human infections ranging from mild non-bloody diarrhea to haemorrhagic colitis (HC), haemolytic uraemic syndrome (HUS), end-stage renal disease (ESRD) and death (Paton and Paton, 1998; Tarr et al., 2005; Gould et al., 2009). Globally, an estimated 2.8 million STEC infections resulted in 3,890 cases of HUS, 270 cases of

ESRD and 230 deaths in 2010 (Majowicz et al., 2014). Importantly, STEC infections were more frequent and severe in children younger than 5 years old (Gould et al., 2009; Buvens et al., 2012; Lozer et al., 2013).

Currently, there are over 470 STEC serotypes recognized based on *E. coli* O (determination of O serogroup) and H (flagellar) antigen typing (Gyles, 2007; Mora et al., 2011; Ludwig et al., 2020). More than 130 STEC serotypes are associated with human STEC infections (Johnson et al., 1996; Bettelheim, 2000; Johnson et al., 2006; Valilis et al., 2018). STEC O157:H7 is the most frequent STEC serotype associated with foodborne outbreaks and human infections (Bettelheim, 2000; Qin et al., 2015; Li et al., 2017). However, other STEC non-O157:H7 serotypes have also been a major cause of foodborne outbreaks and sporadic cases and are responsible for up to 50% STEC infections in recent years (Paton et al., 1999; McCarthy et al., 2001; Paciorek, 2002; Liptáková et al., 2005; Johnson et al., 2006; Zhang et al., 2007; European Food Safety Authority, 2011; Frank et al., 2011a; Käppeli et al., 2011; Verstraete et al., 2013; Zweifel et al., 2013; Morton et al., 2017). Among STEC non-O157:H7 serotypes, 6 serogroups O26, O45, O103, O111, O121 and O45, also known as “Big 6” (comprising 9 serotypes: O26:H11/H-; O45:H2; O103:H2, H11, H25; O111:H8/H-; O121:H19 or H7; and O145:H28/H-) account for over 70% of non-O157:H7 STEC infections (Brooks et al., 2005; Hedican et al., 2009; Bosilevac and Koohmaraie, 2011).

Shiga toxin (Stx) is the main characteristic that defines STEC (Nataro and Kaper, 1998; Tarr et al., 2005), which is encoded by *stx* genes located within lambdoid prophages (Stx-converting phages or Stx-phages) (O'Brien et al., 1989; Mizutani et al., 1999; Bryan et al., 2015; Lacher et al., 2016). Shiga toxins are classified into two types, Stx1 and Stx2. Each of Stx type comprises several subtypes with 3 subtypes for Stx1 (Stx1a, Stx1c and Stx1d) and 10 subtypes for Stx2 (Stx2a, Stx2b, Stx2c, Stx2d, Stx2e, Stx2f, Stx2g, Stx2h, Stx2i and Stx2k) (Scheutz et al., 2012; Lacher et al., 2016; Bai et al., 2018; Yang et al., 2020). Stx1 and/or Stx2 carrying STEC can cause human disease, however, Stx2 is more often associated with HC and HUS (Lentz et al., 2011; Krüger and Lucchesi, 2015). Among Stx2 subtypes, Stx2a is the most prevalent subtype association with severe disease, followed by Stx2c and Stx2d (Feng and Reddy, 2013; Melton-Celsa, 2014; Krüger and Lucchesi, 2015). *Shigella dysenteriae* and some strains of *Shigella sonnei*,



*Shigella flexneri* and *E. albertii* also produce Stx (Beutin et al., 1999; Gupta et al., 2007; Ooka et al., 2012; Gray et al., 2014; Murakami et al., 2014; Brandal et al., 2015). In addition to Shiga toxin, some STEC serotypes also carry the locus of enterocyte effacement (LEE) pathogenicity island (McDaniel and Kaper, 1997; Kaper et al., 2004) responsible for adherence during STEC infections.

STEC detection and identification rely on the detection of Stx proteins by enzyme immune assays or detection of the presence of *stx* genes by molecular methods such as PCR (Brian et al., 1992; Milley and Sekla, 1993; Bélanger et al., 2002; Hara-Kudo et al., 2007; Teel et al., 2007; Zhang et al., 2012). Conventional phenotypic serotyping through antigenic agglutination can further classify STEC to the serotype level (Gyles, 2007). However, cross-reactivity, lack of expression of O antigens, a focus on STEC O157:H7 and novel serotypes may all prevent accurate serotyping and lead to under-detection of STEC non-O157:H7 (Liu et al., 2008; Stigi et al., 2012). Molecular methods, including microarrays, utilising the sequence variations in the O antigen gene clusters, have been developed to serotype STEC O157:H7, “Big 6” STEC non-O157:H7 and other STEC serotypes (DebRoy et al., 2004; Gonzales et al., 2011; Lin et al., 2011; Norman et al., 2012; Iguchi et al., 2015; Ludwig et al., 2020). More recently, WGS based methods have been developed for *in silico* serotyping STEC, which allow phenotypically untypeable isolates be serotyped *in silico* using O antigen and flagellin H antigen genes (Inouye et al., 2014; Joensen et al., 2015).

Alongside STEC serotyping which is useful in outbreak investigation and for prevalence surveillance (FAO/WHO STEC EXPERT GROUP, 2019), other subtyping methods such as pulsed-field gel electrophoresis (PFGE), multiple locus variable-number tandem repeat analysis (MLVA) and multilocus sequence typing (MLST) were also used for STEC outbreak investigations (Gerner-Smidt et al., 2006; Gyles, 2007; Frank et al., 2011b). Recently, WGS based typing and metagenomic sequencing have been shown to have great potential for STEC surveillance and outbreak investigation with high resolution and specificity (Leonard et al., 2015; Parsons et al., 2016) .

STEC serotypes with the same O and H antigens were generally clustered together and share a common ancestor (Ju et al., 2012). A recent phylogenetic analysis on 276 STECs

belonging to 81 serotypes revealed that some STECs formed discrete clades with clustering associated with sequence types and serotypes (González-Escalona and Kase, 2019). This study aimed to i), identify phylogenetic clusters of STEC through large scale examination of publicly available genomes; ii), identify cluster/serotype-specific genes for detection of STEC isolates and for detection and serotyping of most frequent STEC serotypes through comparative genomic analysis of accessory genomes; iii), develop an automated pipeline for STEC *in silico* cluster typing and serotyping from WGS data based on cluster/serotype-specific gene markers combined with *E. coli* O and H antigen genes.

## 5.4 Materials and Methods

### 5.4.1 Identification of STEC isolates from NCBI database

*E. coli* isolates from the NCBI SRA (National Center for Biotechnology Information Sequence Read Archive) in June of 2020 were queried. The keyword “*Escherichia coli*” was used to retrieve SRA accession numbers of *E. coli* isolates. Raw reads were retrieved from ENA (European Nucleotide Archive). The *stx* genes (*stx*<sub>1</sub>, GenBank accession number M19437; *stx*<sub>2</sub> GenBank accession number X07865) and *ipaH* gene (GenBank accession number M32063) were used to screen *E. coli* reads using Salmon v0.13.0 (Patro et al., 2017). Taxonomic classification for *E. coli* was confirmed by Kraken v1.1.1 (Wood and Salzberg, 2014). Isolates that were positive to any of *stx* genes and negative to *ipaH* gene (to eliminate *Shigella* or enteroinvasive *E. coli* [EIEC]) were selected to form the STEC dataset.

A control dataset that represented the sequence types (STs) and ribosomal STs (rSTs) of *stx* negative *E. coli* (“non-STEC”) isolates were constructed. STs and rSTs of non-STEC isolates were obtained from the *E. coli/Shigella* database in the Enterobase on August 2020 (Zhou et al., 2020). For STs and rSTs with only one isolate, the isolate was selected. For STs and rSTs with more than one isolate, one representative isolate for each ST and rST were randomly selected. In total, 14,126 *stx*-negative *E. coli* isolates representing 4,354 STs and 11,520 rSTs were selected as non-STEC control database.

### 5.4.2 Genome assembly and data processing

Raw reads were *de novo* assembled using SPADIS v3.14.0 assembler with default settings [<http://bioinf.spbau.ru/spades>] (Bankevich et al., 2012). The metrics of assembled

genomes were obtained with QUAST v5.0.0 (Gurevich et al., 2013). Three standard deviations (SD) from the mean for contig number, largest contig, total length, GC, N50 and genes were used as quality filter for assembled genomes.

The STs for isolates in the STEC database were checked using mlst (<https://github.com/tseemann/mlst>) with the *E. coli* scheme from PubMLST (Jolley and Maiden, 2010). rSTs were extracted from the *E. coli/Shigella* rMLST database in Enterobase on August 2020 (Zhou et al., 2020). Serotyping of *E. coli* O and H antigen types were predicted by using SerotypeFinder v2.0.1 (Joensen et al., 2015). The phylogroups of STEC isolates were obtained using ClermonTyping (Beghain et al., 2018).

#### **5.4.3 Selection of isolates for STEC identification dataset**

Representative isolates for each ST, rST and serotype in the STEC dataset were selected to form the identification dataset. For STs, rSTs and serotypes with only one isolate, the one isolate was selected. For STs, rSTs and serotypes with more than one isolate, one representative isolate for each ST, rST and serotype was randomly selected. For rSTs in top 6 STs, one representative isolate for each rST with two or more isolates was randomly selected. A further 691 isolates including 72 ECOR isolates downloaded from Enterobase, 573 non-STEC *E. coli* isolates representing 573 STs with more than 9 genomes, 41 *Shigella* and EIEC isolates representing each cluster identified in our previous study (Zhang et al., 2021), 3 *E. albertii* isolates and 2 *E. fergusonii* isolates were used as controls for the identification dataset. The details of the identification dataset are listed in Table S1. The remaining STEC isolates in the STEC database were referred to as the validation dataset (Table S2).

The identification dataset was used to identify the phylogenetic relationships of STEC isolates and was also used to identify cluster/serotype-specific gene markers. The validation dataset was used to evaluate the performance of cluster/serotype-specific gene markers relative to phylogenetic relationships.

#### **5.4.4 Phylogeny of STEC isolates based on WGS**

Phylogenetic trees including an identification tree and 15 validation trees were constructed by using Quicktree v1.3 (Hu et al., 2020) with default parameters to identify

and confirm the phylogenetic clustering of STEC isolates. The phylogenetic trees were visualised by Grapetree and ITOL v5 (Zhou et al., 2018; Letunic and Bork, 2019).

The identification phylogenetic tree was generated using isolates in the identification dataset for the identification of clusters of STEC isolates. The validation trees were constructed using isolates in the STEC validation dataset and a subset of isolates known to represent each identified cluster from the identification dataset to assign validation dataset isolates to the clusters defined.

#### **5.4.5 Identification of the cluster/serotype-specific gene markers**

Cluster/serotype-specific gene markers were identified from STEC accessory genomes. The genomes from the identification dataset were annotated using PROKKA v1.13.3 (Seemann, 2014). Pan- and core-genomes were analysed by Roary v3.12.0 (Page et al., 2015) using an 80% sequence identity threshold. The candidate gene markers specific to each cluster/serotype were identified from accessory genes with an in-house python script from previous study (Zhang et al., 2021). The best performing specific gene marker set was selected from the candidates by using BLASTN to search against the identification dataset.

As in our previous studies (Zhang et al., 2019; Zhang et al., 2021) the genomes from a given cluster containing all specific gene markers for that cluster were termed true positives (TP), the genomes from the same cluster lacking any of those same gene markers were termed false negatives (FN). The genomes from other clusters containing all of those same gene markers were termed false positives (FP). The sensitivity (True positive rate, TPR) of each cluster-specific gene marker was defined as  $TP/(TP+FN)$ . The specificity (True negative rate, TNR) was defined as  $TN/(TN+FP)$ .

#### **5.4.6 Validation of the cluster/serotype-specific gene markers**

The specific gene markers were examined by using BLASTN to search against the validation dataset (Table S2) and non-STEC *E. coli* control database for the presence of any of the cluster/serotype-specific gene markers. The BLASTN thresholds were defined as 80% sequence identity and 50% gene length coverage.

#### **5.4.7 Detection of the cluster/serotype-specific gene markers in STEC spiked food samples using shotgun metagenomic sequencing reads**

The 17 shotgun metagenomic sequencing reads used in Buytaers' study (Buytaers et al., 2020) were downloaded from ENA and trimmed by using Trimmomatic v0.38.0 (Bolger et al., 2014). The detection of cluster/serotype-specific gene markers in the 17 shotgun metagenomic sequencing reads was performed using SRST2 (Inouye et al., 2014).

#### **5.4.8 Development of STECFinder, an automated pipeline for molecular serotyping of STEC**

STECFinder was developed for STEC serotyping from either paired end Illumina genome sequencing reads or assembled genomes. The typing reference sequences used for construction of STECFinder included specific gene marker sets identified in this study, established *E. coli* O antigen and H antigen gene sequences collected from SerotypeFinder (Joensen et al., 2015), *stx* subtypes sequences collected from VirulenceFinder and 3 other studies (Joensen et al., 2014; Lacher et al., 2016; Bai et al., 2018; Yang et al., 2020), *ipaH* gene downloaded from NCBI, and 7 House Keeping (HK) genes *-recA*, *purA*, *mdh*, *icd*, *gyrB*, *fumC* and *adk* from the MLST scheme (Jolley and Maiden, 2010) for contamination checking (Figure 4). All sequences are listed in fasta format available at <https://github.com/LanLab/STECFinder>.

For the submission of sequence data as raw reads, KMA (*k*-mer alignment) v1.3.15 (Clausen et al., 2018) was used to align the raw reads to the typing reference sequences. KMA utilizes *k*-mer seeding and the Needleman-Wunsch algorithm (Needleman and Wunsch, 1970) to accurately align reads to genes of interest. The best-aligning template was chosen from a novel sorting scheme ConClave scheme incorporated into KMA (Clausen et al., 2018). To determine whether the genes were present or absent, the mapping length coverage and a minimum depth were used as the thresholds for determining genes with KMA.

For the submission of sequence data as assembled genomes, BLASTN v2.9.0 (Camacho et al., 2009) was used to search against the typing reference sequences with 80% sequence identity. The presence or absence of genes was determined by the gene length coverage.

The presence or absence of genes in STECFinder was determined by the cutoff value of gene length coverage for assembled genomes and the mapping length coverage and a minimum mapping depth for raw reads. For assembled genomes, length coverage of 50% for all cluster/serotype-specific genes, 60% for O and H antigen genes and 10% for *ipaH* gene and *stx* genes were used as cutoff value for determination of the presence of genes. For raw reads, mapping length coverage of 50% for all cluster/serotype-specific genes, 60% for O and H antigen genes, 10% for *ipaH* gene and *stx* genes and a minimum depth of 10 for all cluster-specific genes, a minimum depth of 1 for O and H antigen genes, *ipaH* gene and *stx* genes were used to define the gene as present. In addition, when multiple O and H genes were detected the bitscore was incorporated into STECFinder for filtering and ranking O and H antigen. The highest match was chosen as the O or H antigen present, when multiple O or H variants were present.

The major and minor clusters and top 10 non-O157:H7 STEC serotypes were assigned based on the presence of cluster/serotype-specific gene marker set together with the presence of *stx* subtypes and the absence of *ipaH* gene. All genes in a cluster/serotype-specific gene set must be defined as present for a cluster or serotype to be called. An 'unclustered' was assigned for isolate that cannot be detected by any of cluster-specific gene marker set. The unclustered STEC could be any new clusters or isolates that contained all genes in the marker set but not all genes from marker set met the cutoff value for presence and therefore classified as unclustered.

Additional subsets of gene marker sets were added to increase the accuracy of clusters and calling of the top 10 non-O157:H7 STEC serotypes. For example, the combination of specific gene marker set of O157:H7 and AM18 can eliminate the known false presences of AM18 gene set in O157:H7. The isolate is assigned as AM18 if both gene sets are present while the isolate is assigned as O157:H7 if AM18 specific gene set is absent. The subsets of combined gene sets incorporated into the STECFinder for elimination of false cluster assignment are listed in Table S6.

STECFinder was tested with identification dataset. The accuracy and specificity of STECFinder for prediction of clusters and serotypes were evaluated with STEC validation dataset and non-STEC *E. coli* control dataset.

## 5.5 Results

### 5.5.1 Screening sequenced genomes for STEC isolates

The presence of any of *stx* genes and the absence of the *ipaH* gene were used to identify STEC isolates. We examined 140,348 isolates with the species annotation of *E. coli* with paired end illumina sequencing reads available in ENA database. Of the 140,348 isolates, 43,960 isolates were positive to *stx*<sub>1</sub> and/or *stx*<sub>2</sub> genes and negative to the *ipaH* gene. 41,101 of the 43,960 isolates passed taxonomic classification and genome assembly quality filters and were selected to form the STEC dataset.

Isolates in the STEC dataset were typed using MLST, rMLST and SerotypeFinder. MLST typed the 41,101 STEC isolates into 817 STs (202 isolates not typed by MLST ) of which 368 STs were represented by a single isolate, 424 STs represented by 2 to 100 isolates each and accounted for 12% of the STEC isolates, whereas 25 STs contained more than 100 isolates each and encompassed 86.61% of the STEC isolates, of which ST11 is the largest, accounting for 37.12% of the STEC isolates, followed by ST21 (14.71%), ST17 (11.91%), ST16 (6.72%), ST655(2.71%) and ST32 (2.46%). rMLST divided the 41,101 STEC isolates into 2,911 rSTs (12,208 isolates not typed by rMLST).

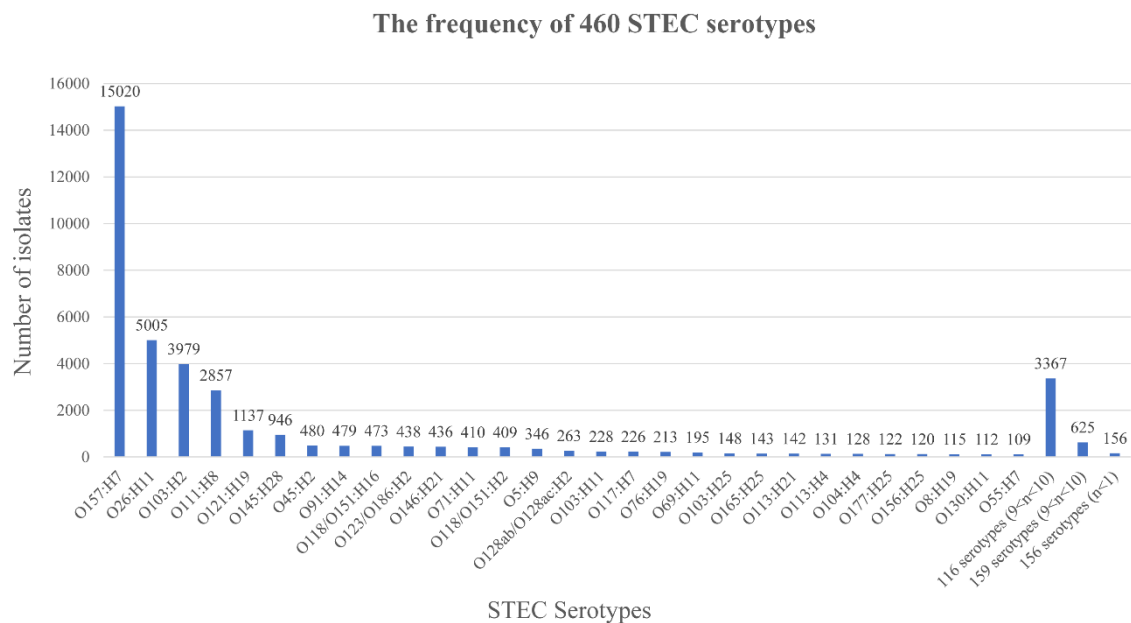
Using SerotypeFinder, 38,958 of the 41,101 (94.79%) isolates were assigned to 460 *E. coli* O:H antigen types, 2,039 isolates (4.96%) were not assigned to O antigen and typed for H antigens only with 38 H antigen types, of which H7, H2, H8, H11 and H21 were the most frequent types, 96 isolates (0.23%) were typed as multiple O:H types and 6 isolates (0.01%) were untypeable.

### 5.5.2 The frequency of STEC serotypes

The 38,958 STEC O:H antigen typeable isolates belonged to 460 different serotypes including O157:H7 (38.55 % of 38,958 typeable isolates) and 459 non-O157:H7 serotypes (61.45% of 38,958 typeable isolates).

Of the 459 non-O157:H7 serotypes, the top 28 serotypes were present in more than 100 isolates each and accounted for 50.8% of 38,958 typeable STEC isolates, of which the 10 most frequent serotypes (41.66% of 38,958 typeable STEC isolates) were O26:H11,

O103:H2, O111:H8, O121:H9, O145:H28, O45:H2, O91:H14, O118/O151:H16, O123/O186:H2 and O146:H21. The top 6 serotypes corresponded to the well-known “Big 6” STEC non-O157:H7 serotypes (Brooks et al., 2005; Hedican et al., 2009; Bosilevac and Koohmaraie, 2011). The 116 serotypes present with 10 to 100 isolates each, belonged to 8.64% of typeable STEC isolates. The remaining 315 serotypes with less than 10 isolates each represented 2% of the typeable STEC isolates (Figure 1).



**Chapter 5. Figure 1: The frequency of 463 STEC serotypes.** The graph shows the frequency of 463 STEC serotypes. STEC O157:H7 and top 28 non-O157:H7 serotypes are listed separately. The number on top of each stacked column refers to the number of isolates for each serotype.

### 5.5.3 Identification of STEC clusters

To identify any phylogenetic clusters containing one or more STEC serotypes from the 41,101 STEC isolates, we selected representative isolates to perform phylogenetic analysis as it was impractical to construct a tree with all isolates. The selection was performed on the basis of ST, rST and serotype of the 41,101 STEC isolates. One isolate was selected to represent each ST, rST and serotype for a total of 2,567 STEC isolates. Note that in the case that STs or rSTs overlapped with serotype, an isolate was only selected once to avoid duplicates of the same isolate. The selection included 817 STs, 1,413 rSTs, 460 STEC serotypes and 102 partial antigen types (H antigen only and multiple O/H types). A further 691 isolates consisting of 72 ECOR isolates, 573 non-



STEC *E. coli* isolates, 41 *Shigella* and EIEC isolates, 3 *E. albertii* isolates and 2 *E. fergusonii* isolates were also included. The identification dataset consisted of 3,258 isolates in total. Details are listed in Table S1. A phylogenetic tree was constructed using 3,258 isolates in the identification dataset to identify the clusters (Figure 2).

The identification of clusters was focused on O157:H7 and top 28 non-O157:H7 serotypes. A major cluster was defined if a branch that only contained STEC isolates and with a bootstrap value of 80% or greater. The isolates of O157:H7 were grouped into one large cluster. A further 18 major clusters (C1-C18) all of which carried only non-O157:H7 serotypes (Figure 2, Table 1, Figure S1), were identified. The isolates of top 28 non-O157:H7 serotypes fell into these 18 major clusters. Of the 2,567 STEC isolates, 1,412 fell within O157:H7 cluster and 18 non-O157:H7 major STEC clusters.

Of the remaining 1,155 STEC isolates, 877 isolates were grouped into 229 STEC minor clusters with 2 or more isolates in a cluster, whereas 278 isolates were singletons separated from other clusters by non-STEC *E. coli* isolates. We further typed the isolates from minor clusters using phylogroup typing (Brooks et al., 2005) and each minor cluster was named by phylogroup and lineage number, for example, phylogroup A minor cluster 1 (AM1). Most of the minor clusters belonged to phylogroup B1 (Table 2).

In total, 19 major STEC clusters including one O157:H7 and 18 non-O157:H7 clusters and 229 STEC minor clusters were identified. Of the 19 major clusters, 12 had a single serotype and 7 had 2 or more serotypes. The frequency of non-O157:H7 STEC serotypes in 19 major clusters are shown in Figure 3. For the 229 STEC minor clusters, 103 contained a single serotype, 109 consisted of 2 or more serotypes and the remaining 17 comprised of isolates with H antigen types only.

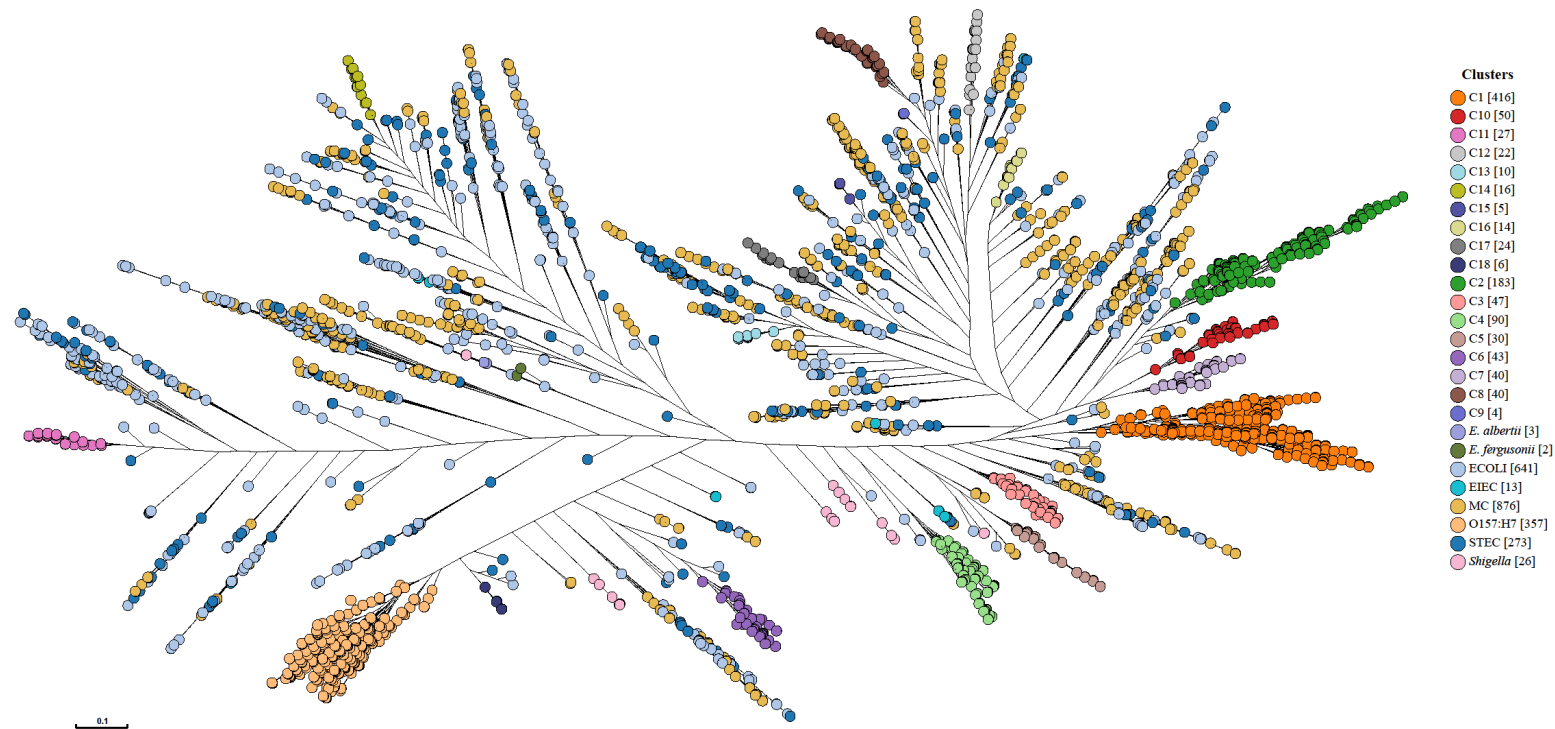
Among the top 10 non-O157:H7 serotypes, O121:H19 (C5), O145:H28 (C6), O91:H14 (C7) had a single origin while O146:H21 (C8 and C9) was a paraphyletic serotype. O26:H11 and O118/O151:H16 were grouped into C1. O123/O186:H2 was grouped into C2. O103:H2, O111:H8 and O45:H2 had polyphyletic origins. O103:H2 and O111:H8 were grouped into C2 and B1M118, C1 and B1M119, respectively. O45:H2 had 3 lineages which were clustered into C2, C3 and AM37. Three serotypes (O128ac:H2,

O8:H19 and O113:H21) of the remaining top 28 non-O157:H7 serotypes were polyphyletic serotypes. Thirty non top 28 non-O157:H7 serotypes also had polyphyletic origins.

**Chapter 5. Table 1: Major STEC clusters identified in identification dataset**

Cluster	No. of isolates	No. of serotypes	No. of STs	Top 28 non-O157:H7 serotypes*
O157:H7	356	1	83	O157:H7
C1	414	30	97	1-O26:H11, 3-O111:H8, 12-O71:H11, 8-O118/O151:H16, 15-O103:H11, 18-O69:H11
C2	181	16	42	2-O103:H2, 6-O45:H2, 9-O123/O186:H2, 11-O118/O151:H2
C3	45	18	12	19-O103:H25, 25-O156:H25, 6-O45:H2
C4	89	14	21	13-O5:H9, 20-O165:H25, 24-O177:H25
C5	29	1	5	4-O121:H19
C6	41	1	6	5-O145:H28
C7	40	2	13	7-O91:H14
C8	40	1	14	10-O146:H21
C9	4	1	1	10-O146:H21
C10	50	2	15	14-O128ab:H2
C11	27	1	6	16-O117:H7
C12	21	1	6	17-O76:H19
C13	10	1	7	21-O113:H21
C14	16	2	2	22-O113:H4
C15	5	1	1	23-O104:H4
C16	14	1	4	26-O8:H19
C17	24	11	7	27-O130:H11
C18	6	1	1	28-O55:H7

\*: The serotypes in each non-O157:H7 cluster is listed with their rank by isolate frequency for the top 28 non-O157:H7 serotypes followed by the serotype.



**Chapter 5. Figure 2: STEC cluster identification phylogenetic tree.** Representative isolates from the identification dataset were used to construct the phylogenetic tree by Quicktree v1.3 to identify STEC clusters and visualised by Grapetree. The tree shows the phylogenetic relationships of 2,567 STEC isolates represented in the identification dataset. Branch lengths are log scale for clarity. The tree scales indicated the 0.1 substitutions per locus. STEC clusters are coloured. Numbers in square brackets indicate the number of isolates for each identified cluster. MC indicates a minor STEC cluster.

Apart from STEC isolates, 26 of the 573 *stx* negative *E. coli* isolates from identification dataset were grouped into clusters. Of the 19 major clusters identified, 12 contained non-STEC *E. coli* isolates (ST11 in O157:H7; ST765 and ST29 in C1; ST17, and ST376 in C2; ST343 and ST300 in C3, ST342 in C4; ST655 in C5; ST32 in C6; ST442 and ST1992 in C8, ST335 in C18). These STs containing *stx* negative *E. coli* isolates were the most frequent STs in the STEC database, suggesting these *stx* negative *E. coli* isolates may have lost the *stx* genes. The details of STEC clusters and lineages were listed in Table S3.

However, 11 STEC minor clusters also contained non-STEC *E. coli* isolates. In this case, it may also be possible that only a subset of isolates within those STs was *stx* positive due to recent acquisition of *stx*. Therefore, we further examined STs with more than 2 isolates from all minor STEC clusters that were also found within the 14,126 *stx* negative *E. coli* (“non-STEC”) isolates. Of the 229 minor STEC clusters, the STs in 58 clusters contained *stx* positive isolates only and the STs in 171 clusters contained both *stx* negative and *stx* positive isolates. Of these 171 minor STEC clusters, the STs in 4 clusters consisted of *stx* positive isolates and *E. coli* isolates that didn’t carry typical pathotype specific genes (data not shown). While STs in the remaining 167 clusters consisted of *stx* positive isolates and the isolates that carried pathotype specific genes from other *E. coli* pathotypes (data not shown). Thus, these STEC minor clusters are a mix of STEC and other pathotypes.

#### **5.5.4 Identification of the cluster/serotype-specific gene markers**

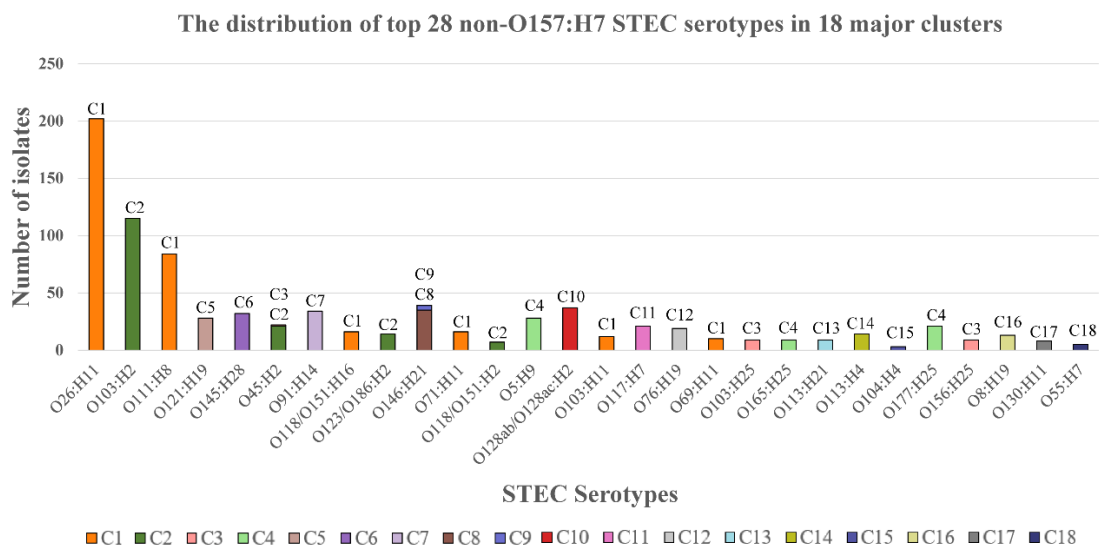
In this study, we used the same definition and approach as used to find the *Shigella*/EIEC cluster specific genes (Zhang et al., 2021). We searched for potential specific gene marker sets for the 19 major and 229 minor clusters using the accessory genomes from the 3,258 identification dataset isolates. Genes associated with STEC O antigen gene clusters were excluded from the analysis to identify O antigen gene independent markers. Multiple candidate cluster/serotype-specific gene marker sets for each of the 19 major STEC clusters and 229 minor STEC clusters were identified. The single gene marker set with 100% sensitive and the highest specificity were then selected from candidate cluster-specific gene marker sets by BLASTN searches against genomes in the identification dataset using 80% sequence identity and 50% gene length threshold.

We also searched for specific gene markers for 6 of the top 10 non-O157:H7 serotypes (O26:H11, O111:H8, O118/O151:H16, O103:H2, O45:H2 and O123/O186:H2) which were not in a cluster of their own. The best performing gene marker set for each of 6 of top 10 non-O157:H7 serotypes were identified using the same approach as used to identify and select cluster-specific gene marker sets.

**Chapter 5. Table 2: Summary of identified STEC minor clusters in identification dataset**

Phylogroup	No. of MC*	Name of MC	No. of isolates	No. of serotypes	No. of STs
A	37	AM1-AM37	139	64	42
B1	126	B1M1-B1M126	519	157	186
B2	14	B2M1-B2M14	35	20	17
C	7	CM1-CM7	17	10	8
D	22	DM1-DM22	67	26	29
E	19	EM1-EM19	73	26	34
G	4	GM1-GM4	27	12	12

\*MC: minor clusters



**Chapter 5. Figure 3: The frequency of STEC serotypes (O157:H7 and top 18 non-O157:H7) in STEC clusters.** The graph shows the frequency of STEC O157:H7 and top 28 non-O157:H7 serotypes in STEC clusters.

We also searched for specific gene markers for 6 of the top 10 non-O157:H7 serotypes (O26:H11, O111:H8, O118/O151:H16, O103:H2, O45:H2 and O123/O186:H2) which were not in a cluster of their own. The best performing gene marker set for each of 6 of top 10 non-O157:H7 serotypes were identified using the same approach as used to identify and select cluster-specific gene marker sets.

The sensitivity and specificity of each major STEC cluster and 6 non-O157:H7 serotype specific gene marker set for the identification dataset were listed in Table 3. The major STEC cluster and 6 non-O157:H7 serotype specific gene marker sets were all 100% sensitive and the specificity varied from 99.72% to 100% for major STEC cluster-specific gene marker set and from 99.41% to 100% for non-O157:H7 serotype-specific gene marker set. The STEC minor cluster-specific gene marker sets were 100% specific with the exception of 12 minor clusters which had specificity ranging from 99.85% to 99.97% (Table S4).

### **5.5.5 Validation of cluster/serotype-specific gene markers**

The STEC cluster/serotype-specific gene marker sets were evaluated with 38,534 STEC isolates from the validation dataset and 14,126 isolates from non-STEC *E. coli* control dataset.

The STEC cluster -specific gene marker sets were able to assign 35,464 of 38,534 (92.03%) STEC isolates to the major clusters and 2,703 (7.01%) STEC isolates to minor clusters. In total, 38,155 of 38,534 (99.02%) STEC isolates can be assigned to clusters by cluster-specific gene marker sets, while 150 of the 38,534 (0.39%) STEC isolates were assigned with more than one cluster and 217 of the 38,534 (0.56%) STEC isolates were not assigned to any cluster by STEC cluster-specific gene marker sets.

Validation phylogenetic trees (Figure S2) were then constructed to confirm the assignment of cluster-specific gene marker sets. We divided the 38,534 STEC validation isolates into 15 subgroups. Each of the 15 subgroups isolates together with a subset of 476 STEC isolates with known clusters and 691 non-STEC isolates from identification dataset were used to generate validation trees for a total of 15 validation trees. The

validation isolates were considered to truly belong to a given cluster if the isolates were found within a branch that only contained identification dataset isolates from that cluster with a bootstrap value of 80% or greater. In total 38,340 (99.5%) validation isolates were assigned to major and minor STEC clusters with 35,574 (92.32%) and 2,766 (7.18%) respectively, while the remaining 194 isolates (0.5%) were not assigned to any clusters.

Compared to cluster assignment by phylogenetic trees as the ground truth, cluster-specific gene marker sets correctly assigned 35,461 validation isolates to major clusters and 2,704 validation isolates to minor clusters. Cluster -specific gene marker sets also correctly identified 191 of the 194 isolates without cluster assignments. In total the accuracy of assignments by cluster -specific gene marker sets were 99.54%. The sensitivity and specificity for each cluster-specific gene marker set for validation dataset were listed in Table S4.

The STEC cluster specific gene marker sets were validated on 14,216 non-STEC *E. coli* isolates. The specificity of the STEC cluster-specific gene markers set for major clusters varied from 99.38% to 100% and the specificity of the STEC cluster-specific gene marker sets for minor clusters ranged from 97.25% to 100%. Details are listed in Table S5.

#### **5.5.6 Detection of the cluster/serotype-specific gene markers in the spiked food samples using shotgun metagenomic sequencing reads**

The application of STEC cluster/serotype-specific gene marker sets in metagenomics analysis was evaluated with 17 metagenomic sequencing reads from samples published by Buytaers *et al.* (Buytaers et al., 2020). The 17 metagenomic samples consisted of 9 minced beef meat samples spiked with a STEC O157:H7 isolate, one fresh goat cheese sample each spiked with STEC O145:H28 isolate, O103:H2 isolate and co-spiked with STEC O103:H2 and O145:H28 isolates and 5 STEC negative control food samples. Samples were spiked with STEC isolates at the lowest infectious dose (<10 CFU for 25 g of food) (Buytaers et al., 2020).

The cluster/serotype-specific gene marker sets were not detected in the 5 control samples. The O157:H7 specific gene set was detected in the expected 9 sequenced reads spiked with STEC O157:H7. The C2 and O103:H2 (O103:H2 is within C2) specific gene sets

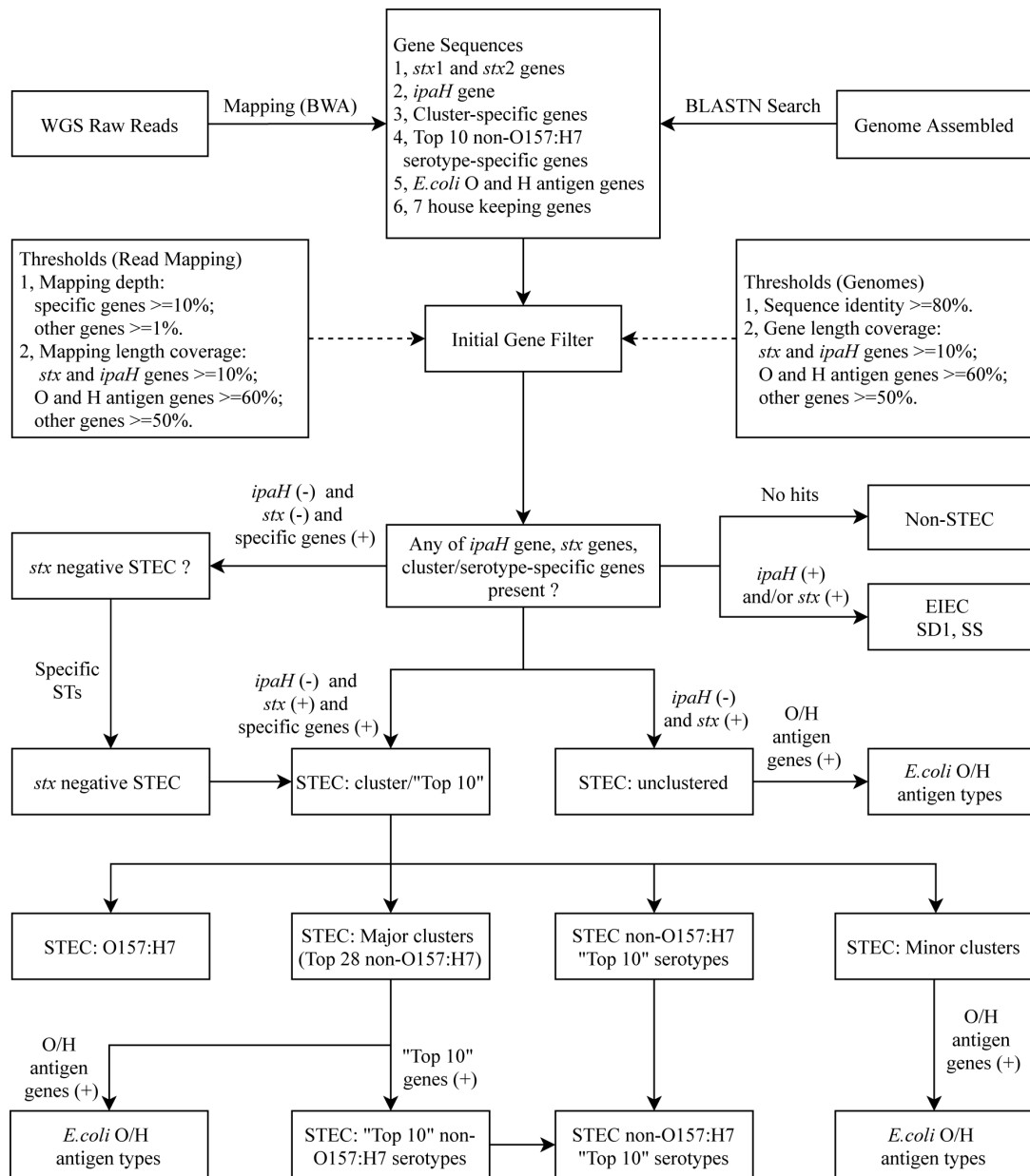
were detected in sequenced reads spiked with STEC O103:H2 and co-spiked with STEC O103:H2 and O145:H28. The C6 (O145:H8) specific gene set was detected in sequenced reads spiked with STEC O145:H28 and co-spiked with STEC O103:H2 and O145:H28.

**Chapter 5. Table 3: The sensitivity and specificity of STEC cluster/serotype-specific gene markers**

Clusters	Cluster-specific gene sets	Identification dataset (3,258 isolates)		
		No of isolates	Sensitivity	Specificity*
O157:H7	Set of 6 genes	356	100	99.72
C1	Set of 4 genes	414	100	99.82
C2	Set of 4 genes	181	100	99.97
C3	Set of 3 genes	45	100	100
C4	Set of 3 genes	89	100	99.97
C5	Set of 4 genes	29	100	100
C6	Set of 3 genes	41	100	99.88
C7	Set of 4 genes	40	100	99.97
C8	Set of 5 genes	40	100	99.97
C9	Set of 2 genes	4	100	100
C10	Set of 2 genes	50	100	100
C11	Single gene	27	100	100
C12	Set of 2 genes	21	100	100
C13	Set of 4 genes	10	100	100
C14	Set of 4 genes	16	100	99.97
C15	Set of 2 genes	5	100	100
C16	Set of 4 genes	14	100	99.97
C17	Set of 3 genes	24	100	99.97
C18	Set of 3 genes	6	100	99.97
O26:H11	Set of 6 genes	204	100	99.41
O103:H2	Set of 4 genes	121	100	99.87
O111:H8	Set of 3 genes	96	100	100
O45:H2 (C2)	Set of 5 genes	22	100	99.97
O45:H2 (C3)	Set of 3 genes	1	100	100
O118/O156:H16	Set of 4 genes	17	100	99.94
O123/O186:H2	Set of 3 genes	21	100	100

\*:The specificity of cluster-specific gene set less than 100% was due to at least one FP found in that set.





**Chapter 5. Figure 4: *in silico* serotyping pipeline workflow.** Schematic of *in silico* serotyping STEC by cluster/serotype-specific genes combined with the *ipaH* gene, *stx* genes including all available subtypes and *E. coli* O antigen and H antigen genes, implemented in STECFinder. Both assembled genomes and raw reads are accepted as data input.

### **5.5.7 STECFinder for molecular serotyping of STEC isolates and its accuracy and specificity**

STECFinder was developed for cluster and serotype identification of STEC isolates. Cluster was identified using cluster -specific gene marker sets and serotype was identified using serotype-specific gene markers as well as *E. coli* O and H antigen genes within clusters. Either paired end Illumina genome sequencing reads or assembled genomes can be used. STECFinder is available on github (<https://github.com/LanLab/STECFinder>).

The accuracy and specificity of STECFinder for STEC typing were tested with 3,258 isolates from the identification dataset. For assembled genomes, all 1,412 STEC isolates belonging to 19 major clusters and all 877 STEC isolates belonging to 229 minor clusters were correctly predicted, while 26 of 573 *stx* negative *E. coli* isolates were assigned to STEC clusters by their corresponding cluster-specific gene marker sets. Eighteen STEC singletons were assigned to clusters or minor clusters. For read mapping, 2 of 1,412 isolates belonging to the 19 major clusters and 25 of 877 isolates from minor clusters were not detected by cluster-specific gene marker sets, while 26 *stx* negative *E. coli* was assigned to STEC clusters similar to the assignment using the assembled genomes. The accuracy of STECFinder for cluster assignments was 99.45% and 98.5% for assembled genomes and read mapping respectively. The accuracy of cluster assignment for the top 10 non-O157:H7 serotypes was 99.14% and 99.11% for assembled genomes and read mapping, respectively.

STECFinder was validated on 38,534 isolates from the STEC validation dataset. Compared to the ground truth assignments determined using phylogenetic analysis, STECFinder assigned 99.85% and 99.83% of validation isolates correctly to clusters for assembled genomes and read mapping, respectively. The accuracy of cluster assignment for top 10 non-O157:H7 serotypes was 99.72% for assembled genomes and 99.65% for read mapping. For the 38,534 *stx*-positive isolates from validation dataset, STECFinder demonstrated 100% cluster assignment specificity for both assembled genomes and read mapping. The cluster assignment specificity of STECFinder was further evaluated using the 14,126 *stx*-negative *E. coli* isolates from the “non-STEC” control dataset. The specificity was 87.07% and 85.12% for assembled genomes and read mapping, respectively. Further investigation of the false positive isolates found that 1,074 false

positive isolates belonged to the STEC cluster based on phylogenetic analysis. After removing all of these false positive isolates, the specificity was 94.66% and 92.72% for assembled genomes and read mapping respectively.

STECFinder can assign STEC isolates to serotype level within predicted clusters. The comparison of *in silico* serotyping of the total of 41,101 STEC isolates between STECFinder and SerotypeFinder (Joensen et al., 2015) was performed. For assembled genomes, the serotype prediction of 40,912 of 41,101 (99.54%) STEC isolates by STECFinder agreed with that by SerotypeFinder when applying the same cutoff values of 80% sequence identity and 60% length coverage. For the remaining 189 STEC isolates with non-identical serotype prediction, STECFinder predicted serotypes were largely a subset of O:H types predicted by SerotypeFinder. For example, an isolate may be assigned as wzx\_O103 and H2 by STECFinder while SerotypeFinder predicted as a mixed wzx\_O103/O26 and H2/H11.

There were 40,618 of 41,101 (98.82%) STEC isolates with the same serotype prediction by STECFinder and SerotypeFinder from read mapping. For the remaining 483 cases, STECFinder assigned a full serotype while SerotypeFinder assigned 257 isolates with H antigen only, 117 and 109 isolates with multiple O:H types.

## 5.6 Discussion

In this study, we performed genomic analysis of nearly 41,000 STEC genomes representing 460 different serotypes and identified 19 major phylogenetic clusters including 1 O157:H7 cluster and 18 non-O157:H7 clusters containing the 28 most frequent non-O157:H7 serotypes, and 229 minor clusters. WGS-based phylogenetic analysis of such a large set of genome data found that STEC had far greater genetic diversity than has been observed previously with clusters containing one or more serotypes. Among the top 28 non-O157:H7 STEC serotypes, 12 serotypes had a single origin. The close phylogenetic relationship between O26:H11, O111:H8 and O103:H11 in C1, O103:H2 and O45:H2 in C2 agreed with previous studies (González-Escalona and Kase, 2019; Zhang et al., 2020). With large number of serotypes (460 serotypes) and polyphyletic and paraphyletic (37 serotypes) origin of many serotypes, identification of serotype specific markers for all serotypes was not possible. However, cluster specific

markers were identified and was used to facilitate accurate prediction and identification of STEC clusters and serotypes. We developed a pipeline STECFinder to facilitate cluster and serotype identification of STEC isolates.

STEC infections have a significant impact on public health worldwide (FAO/WHO STEC EXPERT GROUP, 2019). Early detection and differentiation of STEC is vital for food safety surveillance and public health. The initial screening of *stx* genes for STEC detection may lead to misdiagnosis of STEC because *stx* genes can be lost or transferred (FAO/WHO STEC EXPERT GROUP, 2019). Highly sensitive and specific cluster/serotype-specific gene marker sets identified and evaluated in this study provided robust markers for detection of STEC independent of the presence of *stx* genes. We also identified a small number of *stx*-negative *E. coli* isolates that were grouped into STEC clusters with the corresponding STEC serotypes and STs. Whether these *stx*-negative *E. coli* isolates lost *stx*-containing prophages or were the progenitors of STEC remains unknown. However, human infections caused by *stx*-negative isolates with typical STEC serotypes have been reported previously (Bielaszewska et al., 2007; Mora et al., 2012; Ferdous et al., 2015). STECFinder will predict STEC serotype based on cluster/serotype-specific gene markers even if *stx* is absent.

Our analysis found some minor clusters as well as STs contain both *stx* negative and *stx* positive isolates with *stx* negative isolates being of another *E. coli* pathotypes, which suggests that the STEC within those clusters and STs are hybrid pathogens. These hybrids have been recognised in recent years including the well-known STEC/EAEC (enteroaggregative *E. coli*) hybrid O104:H4 (ST678) and STEC/UPEC (uropathogenic *E. coli*) hybrid of O2:H6 (ST141) (Navarro-Garcia, 2014; Gati et al., 2019). Therefore, for minor STEC clusters, serotypes or STs that carry isolates with different pathogenicity, a note of caution on the use of STECFinder is required as such clusters identified may not uniquely contain STEC pathogens. More data is needed to determine how many serotypes or STs carry different pathotypes.

Serotyping provides valuable information on identification of potential pathogenic STEC (Gyles, 2007; World Health Organization, 2019). Current serotyping methods focus on well-known O157:H7 and “Big 6” non-O157:H7 serotypes. There are many challenges

for detection of other non-O157:H7 serotypes which cause the remaining 20% - 30% STEC infections (DebRoy et al., 2011; Norman et al., 2012; Zweifel et al., 2013; Smith et al., 2014). In addition, not all STEC can be serotyped *in silico* or predicted based on O or H type genes from genome sequencing data (Joensen et al., 2015; González-Escalona and Kase, 2019). The cluster/serotype-specific gene marker sets described here provided nearly perfect prediction of STEC serotypes for the serotypes with 10 or more isolates tested. Cluster-specific gene marker sets and/or serotype-specific gene marker sets can identify O157:H7 and the top 10 most frequent non-O157:H7 serotypes including the “Big 6”. These could be beneficial for identification of the most frequent STEC serotypes for early diagnosis and for clinical management.

Culture-independent approaches such as shotgun metagenomic analysis have been developed for detection of contaminating STEC in enriched food samples as well as mocked food samples (Leonard et al., 2015; Buytaers et al., 2020). We showed that the cluster /serotype-specific gene marker sets of interest were detected in the spiked food samples using shotgun metagenomic sequencing reads used in Buytaers’ study (Buytaers et al., 2020). It is difficult to determine STEC serotype from food or faecal samples directly as O and H antigen genes cannot uniquely identify a STEC serotype in a mixed sample. Our cluster or serotype specific genes provide proxy markers to identify these serotypes in original or non-pure culture samples. These gene marker sets could be adapted for metagenomics based diagnosis and culture independent typing, facilitating rapid STEC identification.

In this study, we developed an automated pipeline STECFinder for *in silico* STEC typing to better inform genomic surveillance of STEC. STECFinder can accurately assign STEC clusters and simultaneously predict *stx* subtypes and STEC serotypes from WGS data. STECFinder can accurately predict all serotypes including those most frequently associated with foodborne outbreaks and severe disease. We verified STECFinder predicted serotype of STEC isolates by phylogenetic cluster assignment and shared STs with STEC isolates of known serotype. Compared with the existing pipeline for *E. coli* *in silico* serotyping, SerotypeFinder (Joensen et al., 2015), cluster/serotype-specific gene markers based STECFinder can eliminate the majority of uncertain antigen calls and provides more accurate STEC O:H typing within predicted clusters. STECFinder will be

useful for epidemiological and diagnostic investigations as well as providing an alternative *in silico* STEC identification method.

We were unable to validate 43 of the 229 minor cluster-specific gene marker sets as these minor clusters had few isolates and once isolates were included in the identification dataset, no isolates remained for validation. Therefore, markers for these 43 minor clusters are tentative and require future validation when more genomes become available. Genes specific to each of these STEC minor clusters were also based on very small number of genomes and should be used with caution. However, since these minor clusters are rarely isolated, they have relatively little effect on the overall applicability of the cluster-specific gene marker sets to STEC typing.

## **5.7 Conclusion**

This study analysed over 41,101 publicly available STEC isolates and identified 19 major and 229 minor STEC clusters. Specific gene marker sets for the 19 major and 229 minor clusters were identified and found to be valuable for *in silico* typing. We also identified serotype specific markers for the top 10 non-O157:H7 STEC serotypes. These markers can be used as proxy markers to identify the serotypes. We additionally developed STECFinder, a freely available *in silico* serotyping pipeline incorporating the cluster/serotype specific gene markers to facilitate serotyping of STEC isolates using genome sequences with very high specificity and sensitivity.

## **5.8 Conflict of Interest**

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## **5.9 Author Contributions**

RL and MP designed the study. XZ, MP and SK performed the bioinformatic analysis. XZ, MP and RL analysed the results. XZ drafted the manuscript. MP and RL provided critical revision of the manuscript.

## 5.10 Abbreviations

STEC, Shiga toxin-producing *Escherichia coli*; HC, haemorrhagic colitis; HUS, haemolytic uraemic syndrome; ESRD, end-stage renal disease; Stx, Shiga toxin; LEE, locus of enterocyte effacement; EIEC, enteroinvasive *E. coli*; MLST, multi-locus sequence typing; rSTs, ribosomal MLST STs; TP, true positives; TPR, true positive rate; TN, true negatives; TNR, true negative rate; FN, false negatives; FP, false positives.

## 5.11 Supplementary tables:

**Table S1:** 3,258 isolates used in identification dataset.

**Table S2:** 38,534 STEC isolates used in validation dataset.

**Table S3:** *stx* negative *E. coli* in STEC clusters and minor clusters in identification dataset.

**Table S4:** The sensitivity and specificity of specific gene marker sets.

**Table S5:** The results of specific gene marker sets tested with 14,126 non-STEC *E. coli* isolates.

**Table S6:** The subsets of cluster/serotype-specific gene markers used in STECFinder.

## 5.12 References

- Bai, X., Fu, S., Zhang, J., Fan, R., Xu, Y., Sun, H., et al. (2018). Identification and pathogenomic analysis of an *Escherichia coli* strain producing a novel Shiga toxin 2 subtype. *Sci Rep* 8(1), 6756. doi: 10.1038/s41598-018-25233-x.
- Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., et al. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* 19(5), 455-477. doi: 10.1089/cmb.2012.0021.
- Beghain, J., Bridier-Nahmias, A., Le Nagard, H., Denamur, E., and Clermont, O. (2018). ClermonTyping: an easy-to-use and accurate *in silico* method for *Escherichia* genus strain phylotyping. *Microb Genom* 4(7). doi: 10.1099/mgen.0.000192.
- Bélanger, S.D., Boissinot, M., Ménard, C., Picard, F.J., and Bergeron, M.G. (2002). Rapid detection of Shiga toxin-producing bacteria in feces by multiplex PCR with molecular beacons on the smart cycler. *J Clin Microbiol* 40(4), 1436-1440. doi: 10.1128/jcm.40.4.1436-1440.2002.
- Bettelheim, K.A. (2000). Role of non-O157 VTEC. *Symp Ser Soc Appl Microbiol* (29), 38s-50s. doi: 10.1111/j.1365-2672.2000.tb05331.x.

- Beutin, L., Strauch, E., and Fischer, I. (1999). Isolation of *Shigella sonnei* lysogenic for a bacteriophage encoding gene for production of Shiga toxin. *Lancet (London, England)* 353(9163), 1498-1498. doi: 10.1016/S0140-6736(99)00961-7.
- Bielaszewska, M., Köck, R., Friedrich, A.W., von Eiff, C., Zimmerhackl, L.B., Karch, H., et al. (2007). Shiga toxin-mediated hemolytic uremic syndrome: time to change the diagnostic paradigm? *PLoS One* 2(10), e1024. doi: 10.1371/journal.pone.0001024.
- Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30(15), 2114-2120. doi: 10.1093/bioinformatics/btu170.
- Bosilevac, J.M., and Koohmaraie, M. (2011). Prevalence and characterization of non-O157 shiga toxin-producing *Escherichia coli* isolates from commercial ground beef in the United States. *Appl Environ Microbiol* 77(6), 2103-2112. doi: 10.1128/aem.02833-10.
- Brandal, L.T., Tunsjø, H.S., Ranheim, T.E., Løbersli, I., Lange, H., and Wester, A.L. (2015). Shiga toxin 2a in *Escherichia albertii*. *Journal of clinical microbiology* 53(4), 1454-1455. doi: 10.1128/jcm.03378-14.
- Brian, M.J., Frosolono, M., Murray, B.E., Miranda, A., Lopez, E.L., Gomez, H.F., et al. (1992). Polymerase chain reaction for diagnosis of enterohemorrhagic *Escherichia coli* infection and hemolytic-uremic syndrome. *J Clin Microbiol* 30(7), 1801-1806. doi: 10.1128/jcm.30.7.1801-1806.1992.
- Brooks, J.T., Sowers, E.G., Wells, J.G., Greene, K.D., Griffin, P.M., Hoekstra, R.M., et al. (2005). Non-O157 Shiga toxin-producing *Escherichia coli* infections in the United States, 1983-2002. *The Journal of infectious diseases* 192(8), 1422-1429. doi: 10.1086/466536.
- Bryan, A., Youngster, I., and McAdam, A.J. (2015). Shiga Toxin Producing *Escherichia coli*. *Clin Lab Med* 35(2), 247-272. doi: 10.1016/j.cl.2015.02.004.
- Buven, G., De Gheldre, Y., Dediste, A., de Moreau, A.I., Mascart, G., Simon, A., et al. (2012). Incidence and virulence determinants of verocytotoxin-producing *Escherichia coli* infections in the Brussels-Capital Region, Belgium, in 2008-2010. *J Clin Microbiol* 50(4), 1336-1345. doi: 10.1128/jcm.05317-11.
- Buytaers, F.E., Saltykova, A., Denayer, S., Verhaegen, B., Vanneste, K., Roosens, N.H.C., et al. (2020). A Practical Method to Implement Strain-Level Metagenomics-Based



- Foodborne Outbreak Investigation and Source Tracking in Routine. *Microorganisms* 8(8). doi: 10.3390/microorganisms8081191.
- Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., et al. (2009). BLAST+: architecture and applications. *BMC Bioinformatics* 10, 421. doi: 10.1186/1471-2105-10-421.
- Clausen, P., Aarestrup, F.M., and Lund, O. (2018). Rapid and precise alignment of raw reads against redundant databases with KMA. *BMC Bioinformatics* 19(1), 307. doi: 10.1186/s12859-018-2336-6.
- DebRoy, C., Roberts, E., Kundrat, J., Davis, M.A., Briggs, C.E., and Fratamico, P.M. (2004). Detection of *Escherichia coli* serogroups O26 and O113 by PCR amplification of the *wzx* and *wzy* genes. *Appl Environ Microbiol* 70(3), 1830-1832. doi: 10.1128/aem.70.3.1830-1832.2004.
- DebRoy, C., Roberts, E., Valadez, A.M., Dudley, E.G., and Cutter, C.N. (2011). Detection of Shiga toxin-producing *Escherichia coli* O26, O45, O103, O111, O113, O121, O145, and O157 serogroups by multiplex polymerase chain reaction of the *wzx* gene of the O-antigen gene cluster. *Foodborne Pathog Dis* 8(5), 651-652. doi: 10.1089/fpd.2010.0769.
- European Food Safety Authority, E.C.f.D.P.C. (2011). The European Union Summary Report on Trends and Sources of Zoonoses, Zoonotic Agents and Food-borne Outbreaks in 2009. *EFSA Journal: European Food Standards Agency* 9(3), 2090. doi: <https://doi.org/10.2903/j.efsa.2011.2090>.
- FAO/WHO STEC EXPERT GROUP (2019). Hazard Identification and Characterization: Criteria for Categorizing Shiga Toxin-Producing *Escherichia coli* on a Risk Basis(†). *Journal of food protection* 82(1), 7-21. doi: 10.4315/0362-028x.Jfp-18-291.
- Feng, P.C., and Reddy, S. (2013). Prevalences of Shiga toxin subtypes and selected other virulence factors among Shiga-toxigenic *Escherichia coli* strains isolated from fresh produce. *Appl Environ Microbiol* 79(22), 6917-6923. doi: 10.1128/aem.02455-13.
- Ferdous, M., Zhou, K., Mellmann, A., Morabito, S., Croughs, P.D., de Boer, R.F., et al. (2015). Is Shiga Toxin-Negative *Escherichia coli* O157:H7 Enteropathogenic or Enterohemorrhagic *Escherichia coli*? Comprehensive Molecular Analysis Using

- Whole-Genome Sequencing. *J Clin Microbiol* 53(11), 3530-3538. doi: 10.1128/jcm.01899-15.
- Frank, C., Faber, M.S., Askar, M., Bernard, H., Fruth, A., Gilsdorf, A., et al. (2011a). Large and ongoing outbreak of haemolytic uraemic syndrome, Germany, May 2011. *Euro Surveill* 16(21).
- Frank, C., Werber, D., Cramer, J.P., Askar, M., Faber, M., an der Heiden, M., et al. (2011b). Epidemic profile of Shiga-toxin-producing *Escherichia coli* O104:H4 outbreak in Germany. *N Engl J Med* 365(19), 1771-1780. doi: 10.1056/NEJMoal106483.
- Gati, N.S., Middendorf-Bauchart, B., Bletz, S., Dobrindt, U., and Mellmann, A. (2019). Origin and Evolution of Hybrid Shiga Toxin-Producing and Uropathogenic *Escherichia coli* Strains of Sequence Type 141. *J Clin Microbiol* 58(1). doi: 10.1128/jcm.01309-19.
- Gerner-Smidt, P., Hise, K., Kincaid, J., Hunter, S., Rolando, S., Hyytiä-Trees, E., et al. (2006). PulseNet USA: a five-year update. *Foodborne Pathog Dis* 3(1), 9-19. doi: 10.1089/fpd.2006.3.9.
- Gonzales, T.K., Kulow, M., Park, D.J., Kaspar, C.W., Anklam, K.S., Pertzborn, K.M., et al. (2011). A high-throughput open-array qPCR gene panel to identify, virulotype, and subtype O157 and non-O157 enterohemorrhagic *Escherichia coli*. *Mol Cell Probes* 25(5-6), 222-230. doi: 10.1016/j.mcp.2011.08.004.
- González-Escalona, N., and Kase, J.A. (2019). Virulence gene profiles and phylogeny of Shiga toxin-positive *Escherichia coli* strains isolated from FDA regulated foods during 2010-2017. *PLoS One* 14(4), e0214620. doi: 10.1371/journal.pone.0214620.
- Gould, L.H., Demma, L., Jones, T.F., Hurd, S., Vugia, D.J., Smith, K., et al. (2009). Hemolytic uremic syndrome and death in persons with *Escherichia coli* O157:H7 infection, foodborne diseases active surveillance network sites, 2000-2006. *Clin Infect Dis* 49(10), 1480-1485. doi: 10.1086/644621.
- Gray, M.D., Lampel, K.A., Strockbine, N.A., Fernandez, R.E., Melton-Celsa, A.R., and Maurelli, A.T. (2014). Clinical isolates of Shiga toxin 1a-producing *Shigella flexneri* with an epidemiological link to recent travel to Hispaniola. *Emerg Infect Dis* 20(10), 1669-1677. doi: 10.3201/eid2010.140292.

- Gupta, S.K., Strockbine, N., Omondi, M., Hise, K., Fair, M.A., and Mintz, E. (2007). Emergence of Shiga toxin 1 genes within *Shigella dysenteriae* type 4 isolates from travelers returning from the Island of Hispaniola. *Am J Trop Med Hyg* 76(6), 1163-1165.
- Gurevich, A., Saveliev, V., Vyahhi, N., and Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. *Bioinformatics* 29(8), 1072-1075. doi: 10.1093/bioinformatics/btt086.
- Gyles, C.L. (2007). Shiga toxin-producing *Escherichia coli*: an overview. *J Anim Sci* 85(13 Suppl), E45-62. doi: 10.2527/jas.2006-508.
- Hara-Kudo, Y., Nemoto, J., Ohtsuka, K., Segawa, Y., Takatori, K., Kojima, T., et al. (2007). Sensitive and rapid detection of Vero toxin-producing *Escherichia coli* using loop-mediated isothermal amplification. *J Med Microbiol* 56(Pt 3), 398-406. doi: 10.1099/jmm.0.46819-0.
- Hedican, E.B., Medus, C., Besser, J.M., Juni, B.A., Koziol, B., Taylor, C., et al. (2009). Characteristics of O157 versus non-O157 Shiga toxin-producing *Escherichia coli* infections in Minnesota, 2000-2006. *Clin Infect Dis* 49(3), 358-364. doi: 10.1086/600302.
- Hu, D., Liu, B., Wang, L., and Reeves, P.R. (2020). Living Trees: High-Quality Reproducible and Reusable Construction of Bacterial Phylogenetic Trees. *Mol Biol Evol* 37(2), 563-575. doi: 10.1093/molbev/msz241.
- Iguchi, A., Iyoda, S., Seto, K., Morita-Ishihara, T., Scheutz, F., and Ohnishi, M. (2015). *Escherichia coli* O-Genotyping PCR: a Comprehensive and Practical Platform for Molecular O Serogrouping. *J Clin Microbiol* 53(8), 2427-2432. doi: 10.1128/jcm.00321-15.
- Inouye, M., Dashnow, H., Raven, L.A., Schultz, M.B., Pope, B.J., Tomita, T., et al. (2014). SRST2: Rapid genomic surveillance for public health and hospital microbiology labs. *Genome Med* 6(11), 90. doi: 10.1186/s13073-014-0090-6.
- Joensen, K.G., Scheutz, F., Lund, O., Hasman, H., Kaas, R.S., Nielsen, E.M., et al. (2014). Real-time whole-genome sequencing for routine typing, surveillance, and outbreak detection of verotoxigenic *Escherichia coli*. *J Clin Microbiol* 52(5), 1501-1510. doi: 10.1128/jcm.03617-13.
- Joensen, K.G., Tetzschner, A.M., Iguchi, A., Aarestrup, F.M., and Scheutz, F. (2015). Rapid and Easy *In Silico* Serotyping of *Escherichia coli* Isolates by Use of Whole-

- Genome Sequencing Data. *Journal of clinical microbiology* 53(8), 2410-2426. doi: 10.1128/jcm.00008-15.
- Johnson, K.E., Thorpe, C.M., and Sears, C.L. (2006). The emerging clinical importance of non-O157 Shiga toxin-producing *Escherichia coli*. *Clin Infect Dis* 43(12), 1587-1595. doi: 10.1086/509573.
- Johnson, R.P., Clarke, R.C., Wilson, J.B., Read, S.C., Rahn, K., Renwick, S.A., et al. (1996). Growing Concerns and Recent Outbreaks Involving Non-O157:H7 Serotypes of Verotoxigenic *Escherichia coli*. *J Food Prot* 59(10), 1112-1122. doi: 10.4315/0362-028x-59.10.1112.
- Jolley, K.A., and Maiden, M.C. (2010). BIGSdb: Scalable analysis of bacterial genome variation at the population level. *BMC Bioinformatics* 11, 595. doi: 10.1186/1471-2105-11-595.
- Ju, W., Cao, G., Rump, L., Strain, E., Luo, Y., Timme, R., et al. (2012). Phylogenetic analysis of non-O157 Shiga toxin-producing *Escherichia coli* strains by whole-genome sequencing. *J Clin Microbiol* 50(12), 4123-4127. doi: 10.1128/jcm.02262-12.
- Kaper, J.B., Nataro, J.P., and Mobley, H.L. (2004). Pathogenic *Escherichia coli*. *Nat Rev Microbiol* 2(2), 123-140. doi: 10.1038/nrmicro818.
- Käppeli, U., Hächler, H., Giezendanner, N., Beutin, L., and Stephan, R. (2011). Human infections with non-O157 Shiga toxin-producing *Escherichia coli*, Switzerland, 2000-2009. *Emerg Infect Dis* 17(2), 180-185. doi: 10.3201/eid1702.100909.
- Krüger, A., and Lucchesi, P.M. (2015). Shiga toxins and stx phages: highly diverse entities. *Microbiology (Reading)* 161(Pt 3), 451-462. doi: 10.1099/mic.0.000003.
- Lacher, D.W., Gangiredla, J., Patel, I., Elkins, C.A., and Feng, P.C. (2016). Use of the *Escherichia coli* Identification Microarray for Characterizing the Health Risks of Shiga Toxin-Producing *Escherichia coli* Isolated from Foods. *J Food Prot* 79(10), 1656-1662. doi: 10.4315/0362-028x.Jfp-16-176.
- Lentz, E.K., Leyva-Illades, D., Lee, M.S., Cherla, R.P., and Tesh, V.L. (2011). Differential response of the human renal proximal tubular epithelial cell line HK-2 to Shiga toxin types 1 and 2. *Infection and immunity* 79(9), 3527-3540. doi: 10.1128/iai.05139-11.
- Leonard, S.R., Mammel, M.K., Lacher, D.W., and Elkins, C.A. (2015). Application of metagenomic sequencing to food safety: detection of Shiga Toxin-producing

- Escherichia coli* on fresh bagged spinach. *Appl Environ Microbiol* 81(23), 8183-8191. doi: 10.1128/aem.02601-15.
- Letunic, I., and Bork, P. (2019). Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res* 47(W1), W256-w259. doi: 10.1093/nar/gkz239.
- Li, B., Liu, H., and Wang, W. (2017). Multiplex real-time PCR assay for detection of *Escherichia coli* O157:H7 and screening for non-O157 Shiga toxin-producing *E. coli*. *BMC Microbiol* 17(1), 215. doi: 10.1186/s12866-017-1123-2.
- Lin, A., Nguyen, L., Lee, T., Clotilde, L.M., Kase, J.A., Son, I., et al. (2011). Rapid O serogroup identification of the ten most clinically relevant STECs by Luminex microbead-based suspension array. *J Microbiol Methods* 87(1), 105-110. doi: 10.1016/j.mimet.2011.07.019.
- Liptáková, A., Siegfried, L., Kmetová, M., Birosová, E., Kotulová, D., Bencátová, A., et al. (2005). Hemolytic uremic syndrome caused by verotoxin-producing *Escherichia coli* O26. Case report. *Folia Microbiol (Praha)* 50(2), 95-98. doi: 10.1007/bf02931454.
- Liu, B., Knirel, Y.A., Feng, L., Perepelov, A.V., Senchenkova, S.N., Wang, Q., et al. (2008). Structure and genetics of *Shigella* O antigens. *FEMS microbiology reviews* 32(4), 627-653. doi: 10.1111/j.1574-6976.2008.00114.x.
- Lozer, D.M., Souza, T.B., Monfardini, M.V., Vicentini, F., Kitagawa, S.S., Scaletsky, I.C., et al. (2013). Genotypic and phenotypic analysis of diarrheagenic *Escherichia coli* strains isolated from Brazilian children living in low socioeconomic level communities. *BMC Infect Dis* 13, 418. doi: 10.1186/1471-2334-13-418.
- Ludwig, J.B., Shi, X., Shridhar, P.B., Roberts, E.L., DebRoy, C., Phebus, R.K., et al. (2020). Multiplex PCR Assays for the Detection of One Hundred and Thirty Seven Serogroups of Shiga Toxin-Producing *Escherichia coli* Associated With Cattle. *Front Cell Infect Microbiol* 10, 378. doi: 10.3389/fcimb.2020.00378.
- Majowicz, S.E., Scallan, E., Jones-Bitton, A., Sargeant, J.M., Stapleton, J., Angulo, F.J., et al. (2014). Global incidence of human Shiga toxin-producing *Escherichia coli* infections and deaths: a systematic review and knowledge synthesis. *Foodborne Pathog Dis* 11(6), 447-455. doi: 10.1089/fpd.2013.1704.

- McCarthy, T.A., Barrett, N.L., Hadler, J.L., Salsbury, B., Howard, R.T., Dingman, D.W., et al. (2001). Hemolytic-Uremic Syndrome and *Escherichia coli* O121 at a Lake in Connecticut, 1999. *Pediatrics* 108(4), E59. doi: 10.1542/peds.108.4.e59.
- McDaniel, T.K., and Kaper, J.B. (1997). A cloned pathogenicity island from enteropathogenic *Escherichia coli* confers the attaching and effacing phenotype on *E. coli* K-12. *Mol Microbiol* 23(2), 399-407. doi: 10.1046/j.1365-2958.1997.2311591.x.
- Melton-Celsa, A.R. (2014). Shiga Toxin (Stx) Classification, Structure, and Function. *Microbiol Spectr* 2(4), Ehec-0024-2013. doi: 10.1128/microbiolspec.EHEC-0024-2013.
- Milley, D.G., and Sekla, L.H. (1993). An enzyme-linked immunosorbent assay-based isolation procedure for verotoxigenic *Escherichia coli*. *Appl Environ Microbiol* 59(12), 4223-4229. doi: 10.1128/aem.59.12.4223-4229.1993.
- Mizutani, S., Nakazono, N., and Sugino, Y. (1999). The so-called chromosomal verotoxin genes are actually carried by defective prophages. *DNA Res* 6(2), 141-143. doi: 10.1093/dnares/6.2.141.
- Mora, A., Herrerra, A., López, C., Dahbi, G., Mamani, R., Pita, J.M., et al. (2011). Characteristics of the Shiga-toxin-producing enteroaggregative *Escherichia coli* O104:H4 German outbreak strain and of STEC strains isolated in Spain. *Int Microbiol* 14(3), 121-141. doi: 10.2436/20.1501.01.142.
- Mora, A., López, C., Dhahi, G., López-Beceiro, A.M., Fidalgo, L.E., Díaz, E.A., et al. (2012). Seropathotypes, Phylogroups, Stx subtypes, and intimin types of wildlife-carried, shiga toxin-producing *escherichia coli* strains with the same characteristics as human-pathogenic isolates. *Applied and environmental microbiology* 78(8), 2578-2585. doi: 10.1128/aem.07520-11.
- Morton, V., Cheng, J.M., Sharma, D., and Kearney, A. (2017). Notes from the Field: An Outbreak of Shiga Toxin-Producing *Escherichia coli* O121 Infections Associated with Flour - Canada, 2016-2017. *MMWR Morb Mortal Wkly Rep* 66(26), 705-706. doi: 10.15585/mmwr.mm6626a6.
- Murakami, K., Etoh, Y., Tanaka, E., Ichihara, S., Horikawa, K., Kawano, K., et al. (2014). Shiga toxin 2f-producing *Escherichia albertii* from a symptomatic human. *Jpn J Infect Dis* 67(3), 204-208. doi: 10.7883/yoken.67.204.

- Nataro, J.P., and Kaper, J.B. (1998). Diarrheagenic *Escherichia coli*. *Clinical microbiology reviews* 11(1), 142-201.
- Navarro-Garcia, F. (2014). *Escherichia coli* O104:H4 Pathogenesis: an Enteroaggregative *E. coli*/Shiga Toxin-Producing *E. coli* Explosive Cocktail of High Virulence. *Microbiol Spectr* 2(6). doi: 10.1128/microbiolspec.EHEC-0008-2013.
- Needleman, S.B., and Wunsch, C.D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* 48(3), 443-453. doi: 10.1016/0022-2836(70)90057-4.
- Norman, K.N., Strockbine, N.A., and Bono, J.L. (2012). Association of nucleotide polymorphisms within the O-antigen gene cluster of *Escherichia coli* O26, O45, O103, O111, O121, and O145 with serogroups and genetic subtypes. *Applied and environmental microbiology* 78(18), 6689-6703. doi: 10.1128/aem.01259-12.
- O'Brien, A.D., Marques, L.R., Kerry, C.F., Newland, J.W., and Holmes, R.K. (1989). Shiga-like toxin converting phage of enterohemorrhagic *Escherichia coli* strain 933. *Microb Pathog* 6(5), 381-390. doi: 10.1016/0882-4010(89)90080-6.
- Ooka, T., Seto, K., Kawano, K., Kobayashi, H., Etoh, Y., Ichihara, S., et al. (2012). Clinical significance of *Escherichia albertii*. *Emerg Infect Dis* 18(3), 488-492. doi: 10.3201/eid1803.111401.
- Paciorek, J. (2002). Virulence properties of *Escherichia coli* faecal strains isolated in Poland from healthy children and strains belonging to serogroups O18, O26, O44, O86, O126 and O127 isolated from children with diarrhoea. *J Med Microbiol* 51(7), 548-571. doi: 10.1099/0022-1317-51-7-548.
- Page, A.J., Cummins, C.A., Hunt, M., Wong, V.K., Reuter, S., Holden, M.T., et al. (2015). Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* 31(22), 3691-3693. doi: 10.1093/bioinformatics/btv421.
- Parsons, B.D., Zelyas, N., Berenger, B.M., and Chui, L. (2016). Detection, Characterization, and Typing of Shiga Toxin-Producing *Escherichia coli*. *Front Microbiol* 7, 478. doi: 10.3389/fmicb.2016.00478.
- Paton, A.W., Woodrow, M.C., Doyle, R.M., Lanser, J.A., and Paton, J.C. (1999). Molecular characterization of a Shiga toxigenic *Escherichia coli* O113:H21 strain lacking eae responsible for a cluster of cases of hemolytic-uremic syndrome. *J Clin Microbiol* 37(10), 3357-3361. doi: 10.1128/jcm.37.10.3357-3361.1999.

- Paton, J.C., and Paton, A.W. (1998). Pathogenesis and diagnosis of Shiga toxin-producing *Escherichia coli* infections. *Clin Microbiol Rev* 11(3), 450-479.
- Patro, R., Duggal, G., Love, M.I., Irizarry, R.A., and Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods* 14(4), 417-419. doi: 10.1038/nmeth.4197.
- Qin, X., Klein, E.J., Galanakis, E., Thomas, A.A., Stapp, J.R., Rich, S., et al. (2015). Real-Time PCR Assay for Detection and Differentiation of Shiga Toxin-Producing *Escherichia coli* from Clinical Samples. *J Clin Microbiol* 53(7), 2148-2153. doi: 10.1128/jcm.00115-15.
- Scheutz, F., Teel, L.D., Beutin, L., Piérard, D., Buvens, G., Karch, H., et al. (2012). Multicenter evaluation of a sequence-based protocol for subtyping Shiga toxins and standardizing Stx nomenclature. *J Clin Microbiol* 50(9), 2951-2963. doi: 10.1128/jcm.00860-12.
- Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 30(14), 2068-2069. doi: 10.1093/bioinformatics/btu153.
- Smith, J.L., Fratamico, P.M., and Gunther, N.W.t. (2014). Shiga toxin-producing *Escherichia coli*. *Adv Appl Microbiol* 86, 145-197. doi: 10.1016/b978-0-12-800262-9.00003-2.
- Stigi, K.A., Macdonald, J.K., Tellez-Marfin, A.A., and Lofy, K.H. (2012). Laboratory practices and incidence of non-O157 shiga toxin-producing *Escherichia coli* infections. *Emerg Infect Dis* 18(3), 477-479. doi: 10.3201/eid1803.111358.
- Tarr, P.I., Gordon, C.A., and Chandler, W.L. (2005). Shiga-toxin-producing *Escherichia coli* and haemolytic uraemic syndrome. *Lancet* 365(9464), 1073-1086. doi: 10.1016/s0140-6736(05)71144-2.
- Teel, L.D., Daly, J.A., Jerris, R.C., Maul, D., Svanas, G., O'Brien, A.D., et al. (2007). Rapid detection of Shiga toxin-producing *Escherichia coli* by optical immunoassay. *J Clin Microbiol* 45(10), 3377-3380. doi: 10.1128/jcm.00837-07.
- Teunis, P.F., Ogden, I.D., and Strachan, N.J. (2008). Hierarchical dose response of *E. coli* O157:H7 from human outbreaks incorporating heterogeneity in exposure. *Epidemiol Infect* 136(6), 761-770. doi: 10.1017/s0950268807008771.
- Tuttle, J., Gomez, T., Doyle, M.P., Wells, J.G., Zhao, T., Tauxe, R.V., et al. (1999). Lessons from a large outbreak of *Escherichia coli* O157:H7 infections: insights



- into the infectious dose and method of widespread contamination of hamburger patties. *Epidemiol Infect* 122(2), 185-192. doi: 10.1017/s0950268898001976.
- Valilis, E., Ramsey, A., Sidiq, S., and DuPont, H.L. (2018). Non-O157 Shiga toxin-producing *Escherichia coli*-A poorly appreciated enteric pathogen: Systematic review. *Int J Infect Dis* 76, 82-87. doi: 10.1016/j.ijid.2018.09.002.
- Verstraete, K., K, D.E.R., S, V.A.N.W., Piérard, D., L, D.E.Z., Herman, L., et al. (2013). Genetic characteristics of Shiga toxin-producing *E. coli* O157, O26, O103, O111 and O145 isolates from humans, food, and cattle in Belgium. *Epidemiol Infect* 141(12), 2503-2515. doi: 10.1017/s0950268813000307.
- Wood, D.E., and Salzberg, S.L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome biology* 15(3), R46. doi: 10.1186/gb-2014-15-3-r46.
- World Health Organization (2019). *Shiga Toxin-producing Escherichia Coli (STEC) and Food: Attribution Characterization and Monitoring*. World Health Organization.
- Yang, X., Bai, X., Zhang, J., Sun, H., Fu, S., Fan, R., et al. (2020). *Escherichia coli* strains producing a novel Shiga toxin 2 subtype circulate in China. *Int J Med Microbiol* 310(1), 151377. doi: 10.1016/j.ijmm.2019.151377.
- Zhang, W., Bielaszewska, M., Bauwens, A., Fruth, A., Mellmann, A., and Karch, H. (2012). Real-time multiplex PCR for detecting Shiga toxin 2-producing *Escherichia coli* O104:H4 in human stools. *J Clin Microbiol* 50(5), 1752-1754. doi: 10.1128/jcm.06817-11.
- Zhang, W., Mellmann, A., Sonntag, A.K., Wieler, L., Bielaszewska, M., Tschäpe, H., et al. (2007). Structural and functional differences between disease-associated genes of enterohaemorrhagic *Escherichia coli* O111. *Int J Med Microbiol* 297(1), 17-26. doi: 10.1016/j.ijmm.2006.10.004.
- Zhang, X., Payne, M., and Lan, R. (2019). *In silico* Identification of Serovar-Specific Genes for *Salmonella* Serotyping. *Front Microbiol* 10, 835. doi: 10.3389/fmicb.2019.00835.
- Zhang, X., Payne, M., Nguyen, T., Kaur, S., and Lan, R. (2021). Cluster-specific gene markers enhance *Shigella* and Enteroinvasive *Escherichia coli* *in silico* serotyping. *bioRxiv*, 2021.2001.2030.428723. doi: 10.1101/2021.01.30.428723.

- Zhang, Y., Liao, Y.T., Sun, X., and Wu, V.C.H. (2020). Is Shiga Toxin-Producing *Escherichia coli* O45 No Longer a Food Safety Threat? The Danger is Still Out There. *Microorganisms* 8(5). doi: 10.3390/microorganisms8050782.
- Zhou, Z., Alikhan, N.F., Mohamed, K., Fan, Y., and Achtman, M. (2020). The EnteroBase user's guide, with case studies on *Salmonella* transmissions, *Yersinia pestis* phylogeny, and *Escherichia* core genomic diversity. *Genome Res* 30(1), 138-152. doi: 10.1101/gr.251678.119.
- Zhou, Z., Alikhan, N.F., Sergeant, M.J., Luhmann, N., Vaz, C., Francisco, A.P., et al. (2018). GrapeTree: visualization of core genomic relationships among 100,000 bacterial pathogens. *Genome Res* 28(9), 1395-1404. doi: 10.1101/gr.232397.117.
- Zweifel, C., Cernela, N., and Stephan, R. (2013). Detection of the emerging Shiga toxin-producing *Escherichia coli* O26:H11/H- sequence type 29 (ST29) clone in human patients and healthy cattle in Switzerland. *Appl Environ Microbiol* 79(17), 5411-5413. doi: 10.1128/aem.01728-13.

## Chapter 6. General Discussion

### 6.1 Key findings and significance of this study

#### 6.1.1 Key findings of this study

##### ***6.1.1.1 Salmonella serovar-specific gene markers identified for most frequent Salmonella serovars***

*Salmonella* is a highly diverse species with over 2,600 serovars. However, a small proportion of *Salmonella* serovars pass down the food production chain to cause severe illness when they contaminate food products. Serovar detection and identification based on antigen encoding genes can be limited by serovars with highly similar O antigen genes and serovars with distinct genetic lineages [20,174,175,195,494]. The ability to detect and distinguish this small proportion of illness causing serovars can be achieved by detection of gene markers specific to a serovar either from genomic data or with laboratory diagnostic methods. Existing molecular methods that utilize specific gene markers or DNA fragments can only distinguish a small number of serovars [177,178,495].

In Chapter 2, 414 candidate serovar-specific and lineage-specific gene markers were identified for 106 *Salmonella* serovars including 24 polyphyletic serovars and the paraphyletic serovar Enteritidis. This is the largest number of serovar-specific gene markers identified to date and covered all of the most common serovars as well as a number of rare serovars. A new approach using the 131 best performing serovar-specific gene markers was designed for molecular *in silico* serotyping of most common *Salmonella* serovars. This approach has an accuracy of 95.3% for *in silico* prediction of the 106 common *Salmonella* serovars from genomic data.

##### ***6.1.1.2 Seven MCDA assays developed for highly sensitive and specific detection and serotyping of five prevalent Salmonella serovars***

The five *Salmonella* serovars: Typhimurium, Enteritidis, Virchow, Saintpaul, and Infantis caused over 85% of human *Salmonella* infections in Australia [43,496-498]. A simple, rapid, sensitive and specific method to detect *Salmonella* and identify these serovars would be useful for public health investigations. The serovar-specific gene markers

identified in Chapter 2 were utilized in Chapter 3 to develop cost-effective laboratory molecular diagnostics assays to detect them.

Seven *Salmonella* serovar-specific gene markers were selected to develop seven laboratory diagnostic MCDA assays for detection of the top five *Salmonella* serovars in Australia. These seven MCDA assays were shown to be highly sensitive and specific (>93.3%) and can type the five *Salmonella* serovars within 8 minutes. In this thesis, MCDA was employed, however any other molecular amplification method, such as PCR can also be used to detect these markers for typing the serovars.

#### ***6.1.1.3 Cluster-specific gene markers identified for differentiation of Shigella and EIEC***

*Shigella* is a major cause of foodborne diarrhoea disease worldwide [209,210]. *Shigella* and EIEC cause human bacillary dysentery and share similar characteristics [216,272,291,296]. As EIEC infections are non-notifiable in nearly all countries, EIEC has frequently been underreported and misidentified as *shigella* [214,302]. Therefore distinguishing these two pathogens is important for clinical, epidemiological and diagnostic investigations. Current genetic markers and *in silico* pipelines may not discriminate between *Shigella* and EIEC in all cases [216,284,302,304-306,326]. However, multiple phylogenetic clusters identified for *Shigella* and EIEC [272,291] could provide high resolution separation of *Shigella* and EIEC if cluster-specific genomic markers were available.

In Chapter 4, 12 previously defined phylogenetic clusters (3 *Shigella* clusters, 5 *Shigella* outliers and 4 EIEC clusters) [272,291] and 5 new clusters consisting of 2 *Shigella* clusters and 3 EIEC clusters were identified by examining over 17,000 publicly available *Shigella* and EIEC genomes. In addition to *Shigella* and EIEC clusters, 53 sporadic EIEC lineages were also described. Cluster-specific gene markers for each cluster and each sporadic EIEC lineage were then identified for differentiation of *Shigella* and EIEC from genomic data with 99.64% accuracy. An *in silico* pipeline, ShigEiFinder was developed based on these cluster-specific gene markers for accurate differentiation, cluster typing and serotyping of *Shigella* and EIEC with 99.38% accuracy.

#### ***6.1.1.4 Cluster/serotype-specific gene markers identified for identification, clustering and serotyping of STEC***

STEC infections have a significant impact on public health worldwide. The well-known STEC O157:H7 is a leading cause of foodborne infections and HUS in humans [499-501]. However, the incidence of STEC non-O157:H7 serotypes associated with foodborne outbreaks and human infections has increased in recent years [431,502-512]. Detection of STEC infection and determination of the serotype of the causative strain are important for accurate diagnosis and detection of outbreaks for public health control. Existing detection and serotyping methods are focused on STEC O157:H7 and “Big 6” non-O157:H7 serotypes [323,329,330,458-468]. Not all common O157:H7 serotypes associated with foodborne outbreaks and severe disease can be detected and predicted *in silico* based on O or H type genes [329,330]. Furthermore, identification of serotype specific gene markers for all serotypes was impractical due to the presence of polyphyletic distributions of many serotypes. Therefore, identification of phylogenetic clusters of STEC through large scale examination of publicly available genomes can improve identification and serotyping of STEC by detection of cluster-specific genomic markers that limit the possible serotype identity determination of an isolate within a phylogenetic cluster.

In Chapter 5, 19 STEC major clusters containing O157:H7 and the top 28 non-O157:H7 as well as 229 STEC minor clusters containing other non-O157:H7 STEC serotypes have been identified through phylogenetic analysis of nearly 41,000 publicly available STEC genomes with 460 different serotypes. Through comparative genomic analysis of STEC accessory genomes, cluster-specific gene markers for STEC clusters and serotype-specific gene markers for the 10 most common STEC non-O157:H7 were then identified for *in silico* typing of STEC with more than 99.54% accuracy. Based on these gene markers, an *in silico* pipeline, STECFinder was developed for genomic identification, clustering and serotyping of STEC and has more than 99.65% accuracy.

#### **6.1.2 Significance of this study**

This thesis has utilised large datasets of publicly available genome sequences and established high quality and representative WGS data subsets. These data were then used for delineation of phylogenies and identification of pathogen type-specific gene markers

for *Salmonella*, *Shigella*, and STEC. The phylogenetic relationships between and within *Salmonella* serovars have been determined and phylogenetic clusters for *Shigella*/EIEC and STEC have been delineated by WGS based phylogenetic analysis in this thesis. The findings provided the most comprehensive view of the diversity and relationships of clusters and lineages of these pathogens.

The pathogen type-specific gene markers for *Salmonella*, *Shigella*, and STEC have been identified in this thesis based on a systematic approach using WGS data combining with comparative genomic analysis of pathogen accessory genomes. The markers discovered in this thesis form a rich resource of genomic markers for development of methods for robust typing of *Salmonella*, *Shigella* and STEC. These pathogen type-specific gene markers could be useful in the development of more cost-effective molecular assays and could be adapted for metagenomics or culture independent typing.

Seven laboratory diagnostic MCDA assays targeting seven *Salmonella* serovar-specific gene markers have been developed in this thesis for detection of five prevalent *Salmonella* serovars with high specificity and high sensitivity. These rapid typing methods have the potential to be used for culture-independent diagnostic testing. If implemented these tests could change the practice of clinical and food production chain testing by allowing rapid identification of these common serovars.

In addition, type-specific gene marker based tools for *in silico* typing of two pathogens, *Shigella*/EIEC and STEC, have been developed and show high accuracy. Pathogen type-specific gene marker typing tools can facilitate rapid detection and identification of *Salmonella*, *Shigella*, EIEC and STEC from genomic data allowing for public health control and prevention of these pathogens, benefiting public health and food safety in Australia and globally.

## **6.2 Establishment of high quality and representative WGS data for identification of pathogen type-specific gene markers**

With a large number of genomes available, a major challenge was to curate the data and select genomes to best represent the diversity of a given pathogen. To cover the genomic

diversity of each of the four pathogens, *Salmonella*, *Shigella*, EIEC or STEC, the available sequenced genomes of the relevant species were screened, and the four pathogens identified, from NCBI database. Representative isolates were selected based on ST and rST in Enterobase which is a publicly available database and hosts both MLST and rMLST typing data for *Salmonella* and *Shigella/E. coli* [23]. The serotypes of *Salmonella*, *Shigella*, EIEC and STEC predicted by the existing *in silico* pipelines were also taken into consideration [193,195,326,329]. Therefore, there were no bias or under representation of sampling for the purpose on the analysis.

Once the isolates were selected, the raw reads were retrieved from ENA (European Nucleotide Archive, <https://www.ebi.ac.uk/ena>) and were *de novo* assembled using SPADIS v3.10.1 assembler [513]. The assembled genomes quality was assessed using QUAST on assembly size, the number of contigs, the largest contig, GC content, the number of gene predicted by glimmer within QUAST [514]. The isolate was replaced by another isolate of the same ST, rST or serotype if the isolate failed the quality control. Therefore, the genomes used in this study were of high quality.

## **6.3 Establishing a systematic approach for identification of pathogen type-specific gene markers**

### **6.3.1 Establishment of a systematic approach**

The major challenge for rapid, sensitive and specific detection *Salmonella*, *Shigella*, EIEC and STEC is to identify highly discriminatory genomic markers [302,309]. In this study we showed that markers reported to be specific and sensitive for identification of Typhimurium and Enteritidis [200-202,208,515,516] did not perform well when applied to the more complete and comprehensive datasets used here. To overcome these issues, this study has established a systematic approach using WGS data combining with comparative genomic analysis of *Salmonella*, *Shigella*, EIEC and STEC accessory genomes to identify pathogen type-specific gene markers.

The approach included phylogenetic analysis of representative isolates, determination of genes presence or absence, identification of candidate pathogen type-specific gene markers, selection of pathogen type-specific gene markers after initial screening,

validation of pathogen type-specific gene markers and development of pathogen type-specific gene markers based typing tools and assays.

Importantly, the approach used quantitative parameters (FP, FN, sensitivity and specificity) to estimate the performance of pathogen type-specific gene markers. Further, the assignments produced by pathogen type-specific gene markers were confirmed by comparison to “ground truth” assignments. These ground truths were determined using phylogenetic analysis of all isolates not used for the initial marker identification and thus are an independent measure of the markers performance. This approach has successfully identified highly sensitive and specific gene markers for *Salmonella*, *Shigella*, EIEC and STEC. This approach would be applicable to other pathogens for the identification of pathogen type-specific gene markers and development of methods for pathogen typing.

### **6.3.2 Pathogen type-specific gene marker sets increase the sensitivity and specificity of typing**

Pathogen type-specific gene markers identified in this study were either a single gene present in all isolates of a cluster and absent in all other isolates or they were a set of 2 or more genes that as a combination were found only in one cluster. A set of genes as a combination was only present in the target cluster although a subset of these genes may be present in other clusters.

To ensure that all clusters of *Shigella*/EIEC and STEC and top 10 most frequent non-O157:H7 STEC serotypes had type-specific gene markers a set of gene markers for clusters or serotypes was considered where a single gene is not available. By considering a set of markers the chance of finding specific sets vastly increased. Additionally, the combination of genes enhanced the accuracy of cluster-specific gene markers as demonstrated by the 100% sensitivity and very high specificity.



## **6.4 WGS based phylogenetic analysis for *Salmonella*, *Shigella* and STEC**

### **6.4.1 *Salmonella* serovar diversity**

WGS based phylogenetic analysis was used to determine the phylogenetic relationships between and within *Salmonella* serovars in Chapter 2. Of 106 most common *Salmonella* serovars investigated there were 81 monophyletic serovars, 24 polyphyletic serovars and one paraphyletic serovar (Enteritidis). These serovars have been reported as polyphyletic/paraphyletic previously [27].

Polyphyletic serovars arise independently from separate ancestors to form separate distinct genetic lineages. Paraphyletic serovars arise from the same ancestor as a monophyletic serovar but a subset of the clade has become a different serovar. Therefore, a combination of lineage-specific gene markers were needed to identify the majority of the polyphyletic serovars or paraphyletic serovars.

### **6.4.2 WGS based analysis identified phylogenetic clusters of *Shigella* and EIEC**

Previous phylogenetic studies based on housekeeping genes indicated that *Shigella* and EIEC isolates consisted of multiple phylogenetic clusters within the broader *E. coli* species [272,291]. In Chapter 4, all 12 previously defined phylogenetic clusters and 5 new clusters of *Shigella* and EIEC were identified through WGS-based phylogenetic analysis of publicly available *Shigella* and EIEC isolates as well as other representatives of the *E. coli* species. WGS-based phylogenetic analysis provided a high resolution method for assigning *Shigella* and EIEC isolates to clusters.

An additional 53 sporadic EIEC types were also identified. These phylogenetic findings demonstrated that EIEC isolates have a greater genetic diversity than has been observed previously and most of the EIEC isolates were clustered more closely to non-enteroinvasive *E. coli* isolates than to major *Shigella* and EIEC clusters. The sporadic isolates belonging to EIEC were confirmed by examination of the presence of the pINV virulence plasmid. The sporadic EIEC isolates may represent recently formed EIEC lineages through acquisition of pINV.

The phylogenetic findings in this work provided a better understanding of the evolution of *Shigella* and EIEC. Providing highly detailed and overwhelming evidence that *Shigella* and EIEC isolates derived independently from multiple distinct lineages of commensal *E. coli* [272,274,291,295].

#### **6.4.3 WGS based analysis identified phylogenetic clusters of STEC**

STEC contains a large number of serotypes. Previous phylogenetic analysis revealed that some STEC serotypes clustered together and formed discrete clades that were associated with specific sequence types [330,517]. In chapter 5, 19 major and 229 minor STEC clusters were identified through WGS-based phylogenetic analysis of nearly 41,000 publicly available STEC genomes representing 460 different serotypes.

The phylogenetic findings indicated that STEC had far greater genetic diversity than has been observed previously with clusters containing one or more serotypes and many serotypes having polyphyletic origins. These phylogenetic clusters facilitate the separation of O157:H7 and top 28 most frequent non-O157:H7 serotypes from other non-O157:H7 serotypes.

A small number of *stx*-negative *E. coli* isolates belonging to the same ST and serotypes as STEC isolates were grouped into STEC clusters (10 major clusters and 11 minor clusters). These *stx*-negative *E. coli* isolates may have lost *stx*-containing prophages or *stx* genes or may be the ancestral isolates of the ST before acquiring Stx-phages. However, *stx*-negative *E. coli* isolates with typical STEC serotypes have been reported to cause human infections, although it is unclear of their pathogenic mechanisms [404-406]. The clustering of these non-STEC *E. coli* with STEC isolates may therefore be of use in detecting pathogenic strains that may otherwise have been overlooked due to their loss of *stx*. However, it is also possible the *stx* negative isolates had a different mode of pathogenicity and thus within the same ST, there are 2 different types of pathogens as hybrid pathogens. These hybrids have been recognised in recent years including the O104:H4 serotype which carried both enteroaggregative and Shiga toxin pathogenicity [356,432,433].

## **6.5 *Salmonella* serovar prediction using serovar-specific gene markers can enhance or replace existing molecular serotyping methods**

An *in silico Salmonella* serovar prediction approach was designed and evaluated in Chapter 2. This *Salmonella* serovar prediction approach has comparable accuracy to other *in silico* serotyping methods and would be useful as diagnosis moves to culture-independent and metagenomic methods.

Compared with existing antigen encoding genes algorithms, SISTR and SeqSero [193,195], gene marker based serovar prediction detects sequence presence or absence of serovar-specific gene markers rather than detecting sequence variation of the antigen genes for H antigens and therefore should perform better than SISTR and SeqSero. The approach is especially useful for cases at low coverage of the genome such as shotgun metagenomic or culture free typing, while SISTR or SeqSero would require sufficient depth and coverage of the genome to accurately detect sequence variation of the H antigen genes.

The unique method developed in this thesis enhances and complements existing *Salmonella* molecular serotyping methods such as SeqSero by using a completely independent measure of relatedness to assign serotypes [193,195].

## **6.6 *Salmonella* serovar-specific gene markers can be used to predict major serovars across the globe**

*Salmonella* serovar-specific gene markers can predict and identify the 106 common *Salmonella* serovars with an accuracy of 95.3%. The accuracy can be further improved if only the major serovars in a given geographic region are considered. The major serovars of highest public health importance for different regions and globally were examined, as different regions have different public health significance on different serovars depending on prevalence [518].

The top 20 serovars found across different continents were collapsed into a combined list of 46 serovars which cause the vast majority of human *Salmonella* infections globally [518]. A subset of serovar-specific gene markers from these 46 serovars when combined with regional frequencies could provide highly specific serovar prediction for locally important serovars. Here regional frequency as addressed in Chapter 2 refers the frequency of a given serovar in a given region. The frequency of the 46 common serovars showed large differences between regions [518]. These differences can be used to adjust the likelihood of a false positive in a given serovar being observed in a given region.

Therefore, the serovar-specific gene markers combined with the prevalence of all major serovars in each continent could be used to design a panel of genes specific for serovars prevalent to a certain region. For example, a panel of seven genes can be used for laboratory based typing of five most frequent serovars in Australia with an error rate less than 2.4% when Australian regional frequencies of serovars in human infections were considered. The false positive Derby strain is a potential false positive of the Typhimurium gene marker STM4494 based on genomic analysis in Chapter 2. However, this potential false positive rate was less than 0.41% in human infections in Australia when the frequency of Derby in Australian human infections was taken into account, which is very rare at less than 1.5%.

## **6.7 *Salmonella* serovar-specific gene markers can be used to develop laboratory detection and serotyping assays**

*Salmonella* is one of the most common causes of foodborne infections worldwide, including in Australia [496]. The major challenge for rapid, sensitive and specific detection *Salmonella* is to find a set of highly discriminatory genomic markers for the common serovars [302,309]. In previous studies [200-202,208,515,516], genomic markers (STM4493, STM4495, STM4497, *typh*, *lygD*, *sdfI*, *safA*, *prot6E*, *spvC*, *sseL* and *sefA*) were used for identification of *Salmonella* serovars. However, the majority of these markers were limited to the detection and identification of Typhimurium and Enteritidis. In this thesis, genomic markers with high specificity were found for all common serovars.

The feasibilities of using a cutting edge molecular assay platform to detect the serovar-specific gene markers identified in this thesis were conducted in Chapter 3 by focusing on a panel of seven genes for the five most common *Salmonella* serovars in Australia. The isothermal amplification technique MCDA [198] was chosen as an assay platform rather than PCR and LAMP because of its higher sensitivity and speed. The existing *invA* MCDA assay for *Salmonella* [197] was included in the assay as a positive control for the species.

Seven accurate and highly sensitive MCDA assays which amplify the seven serovar-specific gene markers can detect and identify Australia five most frequent serovars on pure culture. Their specificity ranged from 93.3% to 100% which reflected the *in silico* typing error rates of the targeted loci. Seven MCDA assays can produce rapid detectable result in as little as 8 minutes. The *invA* MCDA assay was also used as a control to facilitate interpretation of the results of the seven MCDA assays. The assays compared favourably with published Typhimurium LAMP assay targeting gene STM4497 and Enteritidis LAMP assay targeting gene *safA* [200,201], Typhimurium MCDA targeting STM4494 and Enteritidis-clade B MCDA targeting SEN1384 were nearly 62.5% faster and at least 29-fold more sensitive than LAMP assays.

The seven MCDA assays offer rapid, accurate and sensitive detection and identification of *Salmonella* serovars. The performance of the seven MCDA assays warrants further validation on clinically relevant conditions or further tested by the wide community from different geographic regions. Seven MCDA assays will be useful for serotyping of common *Salmonella* serovars in clinical samples as well as food samples once they have been validated. The seven MCDA assays also demonstrated that serovar-specific gene markers can be useful in the development of more cost-effective laboratory molecular diagnostics assays to detect them.

This thesis work showed clear and concise evidence that a unified approach using serovar-specific gene markers and a common detection assay platform can offer a rapid, accurate and sensitive method for serotyping of common *Salmonella* serovars in the era of culture independent diagnostic testing or metagenomic sequencing.

## 6.8 Cluster-specific gene markers enhance *Shigella* and EIEC differentiation and serotyping

Given the relatively poor performance of existing genetic markers for differentiation of *Shigella* from EIEC [216,284,296,302,304-306], cluster-specific gene markers for each phylogenetic cluster that was exclusively composed of *Shigella* or EIEC isolates were identified in Chapter 5.

The cluster-specific gene markers were found to be valuable for highly accurate molecular identification and differentiation of *Shigella* and EIEC independent of the presence of *ipaH* gene. Previously, the *ipaH* gene has been used to differentiate *Shigella* and EIEC from non-enteroinvasive *E. coli* [214,519-521]. However, both *Shigella* and EIEC contain the *ipaH* gene, making differentiation of *Shigella* from EIEC very difficult [214,276,300,302,326]. The cluster-specific gene markers identified in Chapter 4 were specific to *Shigella* and EIEC when evaluated on non-enteroinvasive *E. coli* isolates and are therefore robust markers for the identification of *Shigella* and EIEC.

A new cluster-specific gene marker based *in silico* typing tool, ShigEiFinder was then developed to differentiate *Shigella* isolates from EIEC with 99.74% accuracy. In contrast, the existing tool ShigaTyper [326] differentiated only 47.6% *Shigella* isolates correctly in the same dataset tested. Cluster-specific gene markers used in ShigEiFinder increased the accuracy of *Shigella* and EIEC differentiation and serotyping in comparison to genetic markers *lacY*, *cadA*, *Ss\_methylase* used in ShigaTyper for identification of *Shigella* from EIEC [326]. These markers (*lacY*, *cadA*, *Ss\_methylase*) failed to discriminate between *Shigella* and EIEC when a larger genetic diversity is considered [296,302,309]. Further, *Ss\_methylase* gene was not specific and was found in other *Shigella* serotypes and EIEC [326].

ShigEiFinder is the best platform available so far for accurate differentiation, cluster typing and serotyping of *Shigella* and EIEC. ShigEiFinder can assign isolates to 59 *Shigella* serotypes and 22 EIEC serotypes once the isolate was assigned to a cluster. ShigEiFinder will be useful for clinical, epidemiological and diagnostic investigations.

## 6.9 Cluster/serotype-specific gene markers improve STEC identification, clustering and serotyping

Non-O157:H7 STEC serotypes associated with foodborne outbreaks and human infections have been reported frequently in recent years [431,502-512]. The current detection and serotyping methods for STEC are focusing on O157:H7 and “Big 6” non-O157:H7 STEC serotypes by detection of serotype O antigen and H antigen genes combined with the presence of *stx* genes [323,458-468]. However, these methods are unable to detect and serotype all common STEC non-O157:H7 serotypes [323,458-468] and any methods based on O or H type genes would be error prone or unusable in culture independent typing as O or H antigen genes will be present in the commensal non-STEC *E. coli* co-present in the sample. Furthermore, the identification of STEC relying on the presence of *stx* genes may lead to misdiagnosis of STEC due to the loss of *stx* genes during infection and isolate culture [522].

With the trend towards identification of pathogen specific genomic markers for detection and serotyping of pathogens using genomics, cluster-specific and serotype-specific gene marker were identified for STEC in Chapter 5. Because identification of serotype-specific gene markers for all STEC serotypes was impractical due to the presence of serotypes of polyphyletic origin or natural low frequency of many serotypes.

The STEC cluster/serotype-specific gene markers were specific when evaluated on non-STEC *E. coli* isolates. Therefore they are robust markers for accurate prediction and identification of STEC isolates independent of the presence of *stx* genes. The STEC cluster/serotype-specific gene markers provided nearly perfect prediction of STEC serotypes in many cases without requiring O or H antigen gene characterisation. However, for serotypes that carry isolates with different pathogenicity, a note of caution is required. There are little data how many serotypes or STs that carry different pathotypes.

An existing *E. coli in silico* serotyping pipeline, SerotypeFinder can be used for *in silico* serotyping of STEC by detection of serotype O and H antigen encoding genes [329]. But SerotypeFinder may not uniquely predict a STEC serotype from genomic data as O and H antigen genes can be present in other non-STEC serotypes. In addition, an isolate may

be predicted as multiple O:H types, partial types or untypeable due to novel types or assembly failure. Therefore, not all STEC can be serotyped *in silico* based on O or H type genes from genome sequencing data [329].

The cluster/serotype-specific gene markers based STECFinder developed in Chapter 5 can assign STEC isolates to clusters and identify O157:H7 and the top 10 most frequent non-O157:H7 serotypes including the “Big 6”. STECFinder provides more accurate STEC O:H typing by eliminating the majority of uncertain antigen type calls within predicted clusters in comparison to existing pipeline SerotypeFinder [329]. STECFinder detects the presence or absence of cluster/serotype-specific gene markers and therefore is especially useful for samples with low coverage of the genome such as shotgun metagenomic or culture free typing.

## **6.10 STEC cluster/serotype-specific gene markers can be adapted for metagenomics based diagnosis for rapid STEC identification**

The detection of foodborne STEC relies on culture based methods that are laborious, time-consuming and expensive. Culture-independent approaches such as shotgun metagenomic analysis has the potential for rapid detection of contaminating STEC from food samples in a shorter time period and at a strain-level [484,488]. In a previous study the STEC serotype from these food samples was determined by detection of O and H antigen genes using shotgun metagenomic sequencing reads [484,488]. However, the detection of O and H antigen genes cannot uniquely identify a STEC serotype from food or faecal samples as genes for both antigens genes can be present in other non-STEC serotypes. Therefore, highly sensitive and specific genomic markers are required for metagenomics based methods for the detection of STEC and its serotypes.

In Chapter 5, STEC cluster/serotype-specific gene marker sets of interest were detected in metagenomic sequencing reads from the spiked food samples used in Buytaers’ study [484]. The STEC cluster/serotype-specific gene markers identified and evaluated in this study were specific and sensitive even in the presence of other non-STEC *E. coli* isolates. This thesis work has demonstrated that STEC cluster/serotype-specific gene marker sets



were able to determine STEC serotype in a mixed sample or non-pure culture samples. These gene marker sets as serotype-specific proxy markers could be adapted for culture-independent typing such as shotgun metagenomic sequencing, facilitating rapid STEC identification. This application indicated that pathogen type-specific gene markers identified in this study can be adapted to culture-independent typing approach for rapid foodborne pathogens identification and source tracking of foodborne outbreaks.

## **6.11 Future directions and serotyping of *Salmonella*, *Shigella* and STEC**

This thesis had involved a lot of manual work, such as sorting out genome data from publicly available databases, selection of representative isolates, identification of phylogenetic clusters and selection and testing of genomic markers. Therefore, the bioinformatic approaches used require further development and automation to make them useful for defining new clusters and identifying and validating new genomic markers. The bioinformatic approaches established in this thesis can also be developed for other pathogens where a simple and accurate method for assigning phylogenetically distinct types to isolate genomic data is required.

For *Salmonella*, the seven MCDA assays developed in Chapter 3 were focused on pure culture to demonstrate laboratory detection and serotyping using pathogen type-specific gene markers identified with the previously mentioned phylogenetic approach. The utility of seven MCDA assays in the era of culture independent diagnostic testing and metagenomic sequencing could be supported by further validation in clinically or food industry relevant conditions. The *Salmonella* serovar-specific gene markers could also be used to develop other laboratory detection and serotyping methods, such as PCR and LAMP based methods, for *Salmonella* serovar detection and serotyping.

Like the MCDA assays for the top five *Salmonella* serovars, the cluster and serotype-specific gene markers for *Shigella*, EIEC and STEC could be used to develop laboratory methods for detection, differentiation and serotyping of *Shigella* and EIEC and STEC major serotypes. As in *Salmonella* such methods would have the advantage of only

requiring the detection of the presence of a specific gene not the accurate base calling on full antigen gene sequence.

Two cluster-specific gene marker based *in silico* pipelines, ShigEiFinder in Chapter 4 and STECFinder in Chapter 5, provide the best platform available currently for molecular identification and serotyping of *Shigella*/EIEC and STEC respectively. However, the clusters identified for *Shigella*/EIEC and STEC in this study were based on the genome sequences available in public databases when this study commenced. The isolates representing any new cluster may not be detected by any of cluster-specific gene markers. In addition, for those clusters with low frequency of the serotype or clusters with rare isolates, the cluster-specific gene markers require future validation when more genomes become available.

## 6.12 Conclusion

This thesis has identified highly sensitive and specific pathogen type-specific gene markers for identification and differentiation of serovars of *Salmonella*, clusters and serotypes of *Shigella*, EIEC and STEC using genomics. These specific gene markers have been used to develop genomics based tools for identification of *Salmonella*, *Shigella*, EIEC and STEC clusters and serotypes with high specificity and high sensitivity. These markers could be adapted for metagenomics or culture independent typing and could also be useful in the development of more cost-effective molecular assays. The outcome of this thesis can be applied to rapid typing of respective pathogens in food, clinical and environmental samples and facilitate surveillance of these pathogens for public health control and prevention.

## References

1. Zhao X, Lin CW, Wang J, Oh DH: **Advances in rapid detection methods for foodborne pathogens.** *J Microbiol Biotechnol* 2014, **24**:297-312.
2. Kirk MD, Pires SM, Black RE, Caipo M, Crump JA, Devleesschauwer B, Döpfer D, Fazil A, Fischer-Walker CL, Hald T, et al.: **World Health Organization Estimates of the Global and Regional Disease Burden of 22 Foodborne Bacterial, Protozoal, and Viral Diseases, 2010: A Data Synthesis.** *PLoS medicine* 2015, **12**:e1001921.
3. Havelaar AH, Kirk MD, Torgerson PR, Gibb HJ, Hald T, Lake RJ, Praet N, Bellinger DC, de Silva NR, Gargouri N, et al.: **World Health Organization Global Estimates and Regional Comparisons of the Burden of Foodborne Disease in 2010.** *PLoS medicine* 2015, **12**:e1001923.
4. Andino A, Hanning I: ***Salmonella enterica*: survival, colonization, and virulence differences among serovars.** *ScientificWorldJournal* 2015, **2015**:520179.
5. Foster T: In *Medical Microbiology*. Edited by Baron S: *University of Texas Medical Branch*; 1996.
6. Lamas A, Miranda JM, Regal P, Vázquez B, Franco CM, Cepeda A: **A comprehensive review of non-enterica subspecies of *Salmonella enterica*.** *Microbiol Res* 2018, **206**:60-73.
7. Tindall BJ, Grimont PAD, Garrity GM, Euzéby JP: **Nomenclature and taxonomy of the genus *Salmonella*.** *Int J Syst Evol Microbiol* 2005, **55**:521-524.
8. Brenner FW, Villar RG, Angulo FJ, Tauxe R, Swaminathan B: ***Salmonella* nomenclature.** *J Clin Microbiol* 2000, **38**:2465-2467.
9. Brenner F, McWhorter-Murlin A: **Identification and serotyping of *Salmonella*.** *J Centers for Disease Control Prevention, Atlanta, GA* 1998.
10. Popoff MY, Le Minor Lo: **Antigenic formulas of the *Salmonella* serovars.** 1997.
11. Issenhuth-Jeanjean S, Roggentin P, Mikoleit M, Guibourdenche M, de Pinna E, Nair S, Fields PI, Weill FX: **Supplement 2008-2010 (no. 48) to the White-Kauffmann-Le Minor scheme.** *Res Microbiol* 2014, **165**:526-530.
12. Grimont PA, Weill F-X: **Antigenic formulae of the *Salmonella* serovars.** *WHO collaborating centre for reference research on Salmonella* 2007, **9**:1-166.
13. Ewing WH: **Edwards and Ewing's identification of Enterobacteriaceae.** 1986.

14. Ryan MP, O'Dwyer J, Adley CC: **Evaluation of the Complex Nomenclature of the Clinically and Veterinary Significant Pathogen *Salmonella***. *Biomed Res Int* 2017, **2017**:3782182.
15. Liu B, Knirel YA, Feng L, Perepelov AV, Senchenkova SN, Reeves PR, Wang L: **Structural diversity in *Salmonella* O antigens and its genetic basis**. *FEMS Microbiol Rev* 2014, **38**:56-89.
16. Hong Y, Cunneen MM, Reeves PR: **The Wzx translocases for *Salmonella enterica* O-antigen processing have unexpected serotype specificity**. *Mol Microbiol* 2012, **84**:620-630.
17. Fitzgerald C, Collins M, van Duyn S, Mikoleit M, Brown T, Fields P: **Multiplex, bead-based suspension array for molecular determination of common *Salmonella* serogroups**. *Journal of clinical microbiology* 2007, **45**:3323-3334.
18. Hashimoto Y, Ezaki T, Li N, Yamamoto H: **Molecular cloning of the ViaB region of *Salmonella typhi***. *FEMS Microbiol Lett* 1991, **69**:53-56.
19. Agasan A, Kornblum J, Williams G, Pratt CC, Fleckenstein P, Wong M, Ramon A: **Profile of *Salmonella enterica* subsp. *enterica* (subspecies I) serotype 4,5,12:i:- strains causing food-borne infections in New York City**. *J Clin Microbiol* 2002, **40**:1924-1929.
20. Wattiau P, Boland C, Bertrand S: **Methodologies for *Salmonella enterica* subsp. *enterica* subtyping: gold standards and alternatives**. *Applied and environmental microbiology* 2011, **77**:7877-7885.
21. Chattaway MA, Langridge GC, Wain J: ***Salmonella* nomenclature in the genomic era: a time for change**. *Sci Rep* 2021, **11**:7494.
22. Achtman M, Wain J, Weill FX, Nair S, Zhou Z, Sangal V, Krauland MG, Hale JL, Harbottle H, Uesbeck A, et al.: **Multilocus sequence typing as a replacement for serotyping in *Salmonella enterica***. *PLoS pathogens* 2012, **8**:e1002776.
23. Alikhan NF, Zhou Z, Sergeant MJ, Achtman M: **A genomic overview of the population structure of *Salmonella***. *PLoS genetics* 2018, **14**:e1007261.
24. Beltran P, Plock SA, Smith NH, Whittam TS, Old DC, Selander RK: **Reference collection of strains of the *Salmonella typhimurium* complex from natural populations**. *J Gen Microbiol* 1991, **137**:601-606.

25. Boyd EF, Wang FS, Beltran P, Plock SA, Nelson K, Selander RK: ***Salmonella* reference collection B (SARB): strains of 37 serovars of subspecies I.** *J Gen Microbiol* 1993, **139 Pt 6**:1125-1132.
26. Boyd EF, Wang FS, Whittam TS, Selander RK: **Molecular genetic relationships of the salmonellae.** *Appl Environ Microbiol* 1996, **62**:804-808.
27. Worley J, Meng J, Allard MW, Brown EW, Timme RE: ***Salmonella enterica* Phylogeny Based on Whole-Genome Sequencing Reveals Two New Clades and Novel Patterns of Horizontally Acquired Genetic Elements.** *mBio* 2018, **9**.
28. Mughini-Gras L, Heck M, van Pelt W: **Increase in reptile-associated human salmonellosis and shift toward adulthood in the age groups at risk, the Netherlands, 1985 to 2014.** *Euro Surveill* 2016, **21**.
29. Bertrand S, Rimhanen-Finne R, Weill FX, Rabsch W, Thornton L, Perevoscikovs J, van Pelt W, Heck M: ***Salmonella* infections associated with reptiles: the current situation in Europe.** *Euro Surveill* 2008, **13**.
30. Lee YC, Hung MC, Hung SC, Wang HP, Cho HL, Lai MC, Wang JT: ***Salmonella enterica* subspecies arizonae infection of adult patients in Southern Taiwan: a case series in a non-endemic area and literature review.** *BMC Infect Dis* 2016, **16**:746.
31. Gal-Mor O, Boyle EC, Grassl GA: **Same species, different diseases: how and why typhoidal and non-typhoidal *Salmonella enterica* serovars differ.** *Front Microbiol* 2014, **5**:391.
32. Agbaje M, Begum RH, Oyekunle MA, Ojo OE, Adenubi OT: **Evolution of *Salmonella* nomenclature: a critical note.** *Folia Microbiol (Praha)* 2011, **56**:497-503.
33. Li P, Liu Q, Luo H, Liang K, Yi J, Luo Y, Hu Y, Han Y, Kong Q: **O-Serotype Conversion in *Salmonella* Typhimurium Induces Protective Immune Responses against Invasive Non-Typhoidal *Salmonella* Infections.** *Front Immunol* 2017, **8**:1647.
34. Majowicz SE, Musto J, Scallan E, Angulo FJ, Kirk M, O'Brien SJ, Jones TF, Fazil A, Hoekstra RM: **The global burden of nontyphoidal *Salmonella* gastroenteritis.** *Clin Infect Dis* 2010, **50**:882-889.

35. Ao TT, Feasey NA, Gordon MA, Keddy KH, Angulo FJ, Crump JA: **Global burden of invasive nontyphoidal *Salmonella* disease, 2010(1).** *Emerg Infect Dis* 2015, **21**:941-949.
36. Tapia MD, Tennant SM, Bornstein K, Onwuchekwa U, Tamboura B, Maiga A, Sylla MB, Sissoko S, Kourouma N, Toure A, et al.: **Invasive Nontyphoidal *Salmonella* Infections Among Children in Mali, 2002-2014: Microbiological and Epidemiologic Features Guide Vaccine Development.** *Clin Infect Dis* 2015, **61 Suppl 4**:S332-338.
37. Akullian A, Montgomery JM, John-Stewart G, Miller SI, Hayden HS, Radey MC, Hager KR, Verani JR, Ochieng JB, Juma J, et al.: **Multi-drug resistant non-typhoidal *Salmonella* associated with invasive disease in western Kenya.** *PLoS Negl Trop Dis* 2018, **12**:e0006156.
38. Jaffee S, Henson S, Unnevehr L, Grace D, Cassou E, Havelaar A, Kirk M, Torgerson P, Gibb H, Hald T: **World Health Organization Global Estimates and Regional Comparisons of the Burden of Foodborne Disease in 2010.** Edited by: *University of Southern California Los Angeles*; 2018.
39. Phu Huong Lan N, Le Thi Phuong T, Nguyen Huu H, Thuy L, Mather AE, Park SE, Marks F, Thwaites GE, Van Vinh Chau N, Thompson CN, et al.: **Invasive Non-typhoidal *Salmonella* Infections in Asia: Clinical Observations, Disease Outcome and Dominant Serovars from an Infectious Disease Hospital in Vietnam.** *PLoS Negl Trop Dis* 2016, **10**:e0004857.
40. Collaborators GTaP: **The global burden of typhoid and paratyphoid fevers: a systematic analysis for the Global Burden of Disease Study 2017.** *Lancet Infect Dis* 2019, **19**:369-381.
41. Saha S, Islam MS, Sajib MSI, Saha S, Uddin MJ, Hooda Y, Hasan M, Amin MR, Hanif M, Shahidullah M, et al.: **Epidemiology of Typhoid and Paratyphoid: Implications for Vaccine Policy.** *Clin Infect Dis* 2019, **68**:S117-s123.
42. Kirk M, Ford L, Glass K, Hall G: **Foodborne illness, Australia, circa 2000 and circa 2010.** *Emerging infectious diseases* 2014, **20**:1857-1864.
43. Group. OW: **Monitoring the incidence and causes of diseases potentially transmitted by food in Australia: Annual report of the OzFoodNet network, 2011.** *Communicable diseases intelligence quarterly report* 2015, **39**:E236.

44. Ford L, Glass K, Veitch M, Wardell R, Polkinghorne B, Dobbins T, Lal A, Kirk MD: **Increasing Incidence of *Salmonella* in Australia, 2000-2013.** *PLoS One* 2016, **11**:e0163989.
45. Cheng RA, Eade CR, Wiedmann M: **Embracing Diversity: Differences in Virulence Mechanisms, Disease Severity, and Host Adaptations Contribute to the Success of Nontyphoidal *Salmonella* as a Foodborne Pathogen.** *Front Microbiol* 2019, **10**:1368.
46. Marder Mph EP, Griffin PM, Cieslak PR, Dunn J, Hurd S, Jervis R, Lathrop S, Muse A, Ryan P, Smith K, et al.: **Preliminary Incidence and Trends of Infections with Pathogens Transmitted Commonly Through Food - Foodborne Diseases Active Surveillance Network, 10 U.S. Sites, 2006-2017.** *MMWR Morb Mortal Wkly Rep* 2018, **67**:324-328.
47. Scallan E, Hoekstra RM, Angulo FJ, Tauxe RV, Widdowson MA, Roy SL, Jones JL, Griffin PM: **Foodborne illness acquired in the United States--major pathogens.** *Emerg Infect Dis* 2011, **17**:7-15.
48. Authority. EFS: **The European Union One Health 2019 Zoonoses Report.** *EFSA J* 2021, **19**:e06406.
49. ECDC) EFSAaECfDPaCEa: **The European Union summary report on trends and sources of zoonoses, zoonotic agents and food-borne outbreaks in 2017.** *EFSA J* 2018, **16**:e05500.
50. Rodriguez A, Pangloli P, Richards HA, Mount JR, Draughon FA: **Prevalence of *Salmonella* in diverse environmental farm samples.** *J Food Prot* 2006, **69**:2576-2580.
51. Ferrari RG, Rosario DKA, Cunha-Neto A, Mano SB, Figueiredo EES, Conte-Junior CA: **Worldwide Epidemiology of *Salmonella* Serovars in Animal-Based Foods: a Meta-analysis.** *Appl Environ Microbiol* 2019, **85**.
52. Organization. WH: **Interventions for the control of non-typhoidal *Salmonella* spp. in beef and pork: meeting report and systematic review.** *World Health Organization* 2016.
53. Authority. EFS: **A quantitative microbiological risk assessment on *Salmonella* in meat: Source attribution for human salmonellosis from meat-Scientific Opinion of the Panel on Biological Hazards.** *The EFSA Journal* 2008, **625**:1-32.

54. Cummings KJ, Rodriguez-Rivera LD, Mitchell KJ, Hoelzer K, Wiedmann M, McDonough PL, Altier C, Warnick LD, Perkins GA: ***Salmonella enterica* serovar Oranienburg outbreak in a veterinary medical teaching hospital with evidence of nosocomial and on-farm transmission.** *Vector Borne Zoonotic Dis* 2014, **14**:496-502.
55. Kariuki S, Revathi G, Kariuki N, Kiiru J, Mwituria J, Muyodi J, Githinji JW, Kagendo D, Munyalo A, Hart CA: **Invasive multidrug-resistant non-typhoidal *Salmonella* infections in Africa: zoonotic or anthroponotic transmission?** *J Med Microbiol* 2006, **55**:585-591.
56. Connor BA, Schwartz E: **Typhoid and paratyphoid fever in travellers.** *Lancet Infect Dis* 2005, **5**:623-628.
57. van Asten AJ, van Dijk JE: **Distribution of "classic" virulence factors among *Salmonella* spp.** *FEMS Immunol Med Microbiol* 2005, **44**:251-259.
58. Ehrbar K, Hardt WD: **Bacteriophage-encoded type III effectors in *Salmonella enterica* subspecies 1 serovar Typhimurium.** *Infect Genet Evol* 2005, **5**:1-9.
59. Coburn B, Sekirov I, Finlay BB: **Type III secretion systems and disease.** *Clin Microbiol Rev* 2007, **20**:535-549.
60. Ghosh P: **Process of protein transport by the type III secretion system.** *Microbiol Mol Biol Rev* 2004, **68**:771-795.
61. Mirolid S, Rabsch W, Tschäpe H, Hardt WD: **Transfer of the *Salmonella* type III effector *sopE* between unrelated phage families.** *J Mol Biol* 2001, **312**:7-16.
62. Figueira R, Holden DW: **Functions of the *Salmonella* pathogenicity island 2 (SPI-2) type III secretion system effectors.** *Microbiology (Reading)* 2012, **158**:1147-1161.
63. Wallis TS, Galyov EE: **Molecular basis of *Salmonella*-induced enteritis.** *Mol Microbiol* 2000, **36**:997-1005.
64. Galán JE, Collmer A: **Type III secretion machines: bacterial devices for protein delivery into host cells.** *Science* 1999, **284**:1322-1328.
65. Beveridge TJ: **Structures of gram-negative cell walls and their derived membrane vesicles.** *J Bacteriol* 1999, **181**:4725-4733.
66. Deatherage BL, Lara JC, Bergsbaken T, Rassouljian Barrett SL, Lara S, Cookson BT: **Biogenesis of bacterial membrane vesicles.** *Mol Microbiol* 2009, **72**:1395-1407.



67. Yoon H, Ansong C, Adkins JN, Heffron F: **Discovery of *Salmonella* virulence factors translocated via outer membrane vesicles to murine macrophages.** *Infect Immun* 2011, **79**:2182-2192.
68. Wai SN, Lindmark B, Söderblom T, Takade A, Westermark M, Oscarsson J, Jass J, Richter-Dahlfors A, Mizunoe Y, Uhlin BE: **Vesicle-mediated export and assembly of pore-forming oligomers of the enterobacterial ClyA cytotoxin.** *Cell* 2003, **115**:25-35.
69. de Jong HK, Parry CM, van der Poll T, Wiersinga WJ: **Host-pathogen interaction in invasive Salmonellosis.** *PLoS Pathog* 2012, **8**:e1002933.
70. Murdoch SL, Trunk K, English G, Fritsch MJ, Pourkarimi E, Coulthurst SJ: **The opportunistic pathogen *Serratia marcescens* utilizes type VI secretion to target bacterial competitors.** *J Bacteriol* 2011, **193**:6057-6069.
71. Schwarz S, Hood RD, Mougous JD: **What is type VI secretion doing in all those bugs?** *Trends Microbiol* 2010, **18**:531-537.
72. Pezoa D, Blondel CJ, Silva CA, Yang HJ, Andrews-Polymenis H, Santiviago CA, Contreras I: **Only one of the two type VI secretion systems encoded in the *Salmonella enterica* serotype Dublin genome is involved in colonization of the avian and murine hosts.** *Vet Res* 2014, **45**:2.
73. Blondel CJ, Yang HJ, Castro B, Chiang S, Toro CS, Zaldívar M, Contreras I, Andrews-Polymenis HL, Santiviago CA: **Contribution of the type VI secretion system encoded in SPI-19 to chicken colonization by *Salmonella enterica* serotypes Gallinarum and Enteritidis.** *PLoS One* 2010, **5**:e11724.
74. Reitmeyer JC, Peterson JW, Wilson KJ: ***Salmonella* cytotoxin: a component of the bacterial outer membrane.** *Microb Pathog* 1986, **1**:503-510.
75. den Bakker HC, Moreno Switt AI, Govoni G, Cummings CA, Ranieri ML, Degoricija L, Hoelzer K, Rodriguez-Rivera LD, Brown S, Bolchacova E, et al.: **Genome sequencing reveals diversification of virulence factor content and possible host adaptation in distinct subpopulations of *Salmonella enterica*.** *BMC Genomics* 2011, **12**:425.
76. Lara-Tejero M, Galán JE: **Cytolethal distending toxin: limited damage as a strategy to modulate cellular functions.** *Trends Microbiol* 2002, **10**:147-152.
77. Saitoh M, Tanaka K, Nishimori K, Makino SI, Kanno T, Ishihara R, Hatama S, Kitano R, Kishima M, Sameshima T, et al.: **The *artAB* genes encode a putative**

- ADP-ribosyltransferase toxin homologue associated with *Salmonella enterica* serovar Typhimurium DT104.** *Microbiology (Reading)* 2005, **151**:3089-3096.
78. Rodriguez-Rivera LD, Bowen BM, den Bakker HC, Duhamel GE, Wiedmann M: **Characterization of the cytolethal distending toxin (typhoid toxin) in non-typhoidal *Salmonella* serovars.** *Gut Pathog* 2015, **7**:19.
79. Hiley L, Fang NX, Micalizzi GR, Bates J: **Distribution of Gifsy-3 and of variants of ST64B and Gifsy-1 prophages amongst *Salmonella enterica* Serovar Typhimurium isolates: evidence that combinations of prophages promote clonality.** *PLoS One* 2014, **9**:e86203.
80. Cheng RA, Wiedmann M: **The ADP-Ribosylating Toxins of *Salmonella*.** *Toxins (Basel)* 2019, **11**.
81. Prager R, Fruth A, Tschäpe H: ***Salmonella* enterotoxin (stn) gene is prevalent among strains of *Salmonella enterica*, but not among *Salmonella bongori* and other Enterobacteriaceae.** *FEMS Immunol Med Microbiol* 1995, **12**:47-50.
82. Lee K, Iwata T, Shimizu M, Taniguchi T, Nakadai A, Hirota Y, Hayashidani H: **A novel multiplex PCR assay for *Salmonella* subspecies identification.** *J Appl Microbiol* 2009, **107**:805-811.
83. Chopra AK, Huang JH, Xu X, Burden K, Niesel DW, Rosenbaum MW, Popov VL, Peterson JW: **Role of *Salmonella* enterotoxin in overall virulence of the organism.** *Microb Pathog* 1999, **27**:155-171.
84. Wallis TS, Wood M, Watson P, Paulin S, Jones M, Galyov E: **Sips, Sops, and SPIs but not stn influence *Salmonella* enteropathogenesis.** *Adv Exp Med Biol* 1999, **473**:275-280.
85. Watson PR, Paulin SM, Bland AP, Jones PW, Wallis TS: **Characterization of intestinal invasion by *Salmonella typhimurium* and *Salmonella dublin* and effect of a mutation in the *invH* gene.** *Infect Immun* 1995, **63**:2743-2754.
86. Smyth CJ, Marron MB, Twohig JM, Smith SG: **Fimbrial adhesins: similarities and variations in structure and biogenesis.** *FEMS Immunol Med Microbiol* 1996, **16**:127-139.
87. Fernández LA, Berenguer J: **Secretion and assembly of regular surface structures in Gram-negative bacteria.** *FEMS Microbiol Rev* 2000, **24**:21-44.

88. Yue M, Rankin SC, Blanchet RT, Nulton JD, Edwards RA, Schifferli DM:  
**Diversification of the *Salmonella* fimbriae: a model of macro- and microevolution.** *PLoS One* 2012, **7**:e38596.
89. Morgan DG, Owen C, Melanson LA, DeRosier DJ: **Structure of bacterial flagellar filaments at 11 Å resolution: packing of the alpha-helices.** *J Mol Biol* 1995, **249**:88-110.
90. Dos Santos AMP, Ferrari RG, Conte-Junior CA: **Virulence Factors in *Salmonella* Typhimurium: The Sagacity of a Bacterium.** *Curr Microbiol* 2019, **76**:762-773.
91. Spöring I, Felgner S, Preuße M, Eckweiler D, Rohde M, Häussler S, Weiss S, Erhardt M: **Regulation of Flagellum Biosynthesis in Response to Cell Envelope Stress in *Salmonella enterica* Serovar Typhimurium.** *mBio* 2018, **9**.
92. McQuiston JR, Parrenas R, Ortiz-Rivera M, Gheesling L, Brenner F, Fields PI: **Sequencing and comparative analysis of flagellin genes *fliC*, *fliB*, and *fliA* from *Salmonella*.** *Journal of clinical microbiology* 2004, **42**:1923-1932.
93. Bonifield HR, Hughes KT: **Flagellar phase variation in *Salmonella enterica* is mediated by a posttranscriptional control mechanism.** *J Bacteriol* 2003, **185**:3567-3574.
94. Hensel M: **Evolution of pathogenicity islands of *Salmonella enterica*.** *Int J Med Microbiol* 2004, **294**:95-102.
95. Hayward MR, AbuOun M, La Ragione RM, Tchórzewska MA, Cooley WA, Everest DJ, Petrovska L, Jansen VA, Woodward MJ: **SPI-23 of *S. Derby*: role in adherence and invasion of porcine tissues.** *PLoS One* 2014, **9**:e107857.
96. Hayward MR, Jansen V, Woodward MJ: **Comparative genomics of *Salmonella enterica* serovars Derby and Mbandaka, two prevalent serovars associated with different livestock species in the UK.** *BMC Genomics* 2013, **14**:365.
97. Sabbagh SC, Forest CG, Lepage C, Leclerc JM, Daigle F: **So similar, yet so different: uncovering distinctive features in the genomes of *Salmonella enterica* serovars Typhimurium and Typhi.** *FEMS Microbiol Lett* 2010, **305**:1-13.
98. Blondel CJ, Jiménez JC, Contreras I, Santiviago CA: **Comparative genomic analysis uncovers 3 novel loci encoding type six secretion systems**

- differentially distributed in *Salmonella* serotypes.** *BMC Genomics* 2009, **10**:354.
99. Fookes M, Schroeder GN, Langridge GC, Blondel CJ, Mammina C, Connor TR, Seth-Smith H, Vernikos GS, Robinson KS, Sanders M, et al.: ***Salmonella bongori* provides insights into the evolution of the Salmonellae.** *PLoS Pathog* 2011, **7**:e1002191.
  100. Rychlik I, Gregorova D, Hradecka H: **Distribution and function of plasmids in *Salmonella enterica*.** *Vet Microbiol* 2006, **112**:1-10.
  101. Liu WQ, Feng Y, Wang Y, Zou QH, Chen F, Guo JT, Peng YH, Jin Y, Li YG, Hu SN, et al.: ***Salmonella paratyphi C*: genetic divergence from *Salmonella choleraesuis* and pathogenic convergence with *Salmonella typhi*.** *PLoS One* 2009, **4**:e4510.
  102. Silva C, Puente JL, Calva E: ***Salmonella* virulence plasmid: pathogenesis and ecology.** *Pathog Dis* 2017.
  103. Akiba M, Sameshima T, Anzai T, Wada R, Nakazawa M: ***Salmonella Abortusequi* strains of equine origin harbor a 95kb plasmid responsible for virulence in mice.** *Vet Microbiol* 1999, **68**:265-272.
  104. Uzzau S, Gulig PA, Paglietti B, Leori G, Stocker BA, Rubino S: **Role of the *Salmonella abortusovis* virulence plasmid in the infection of BALB/c mice.** *FEMS Microbiol Lett* 2000, **188**:15-18.
  105. Chu C, Hong SF, Tsai C, Lin WS, Liu TP, Ou JT: **Comparative physical and genetic maps of the virulence plasmids of *Salmonella enterica* serovars typhimurium, enteritidis, choleraesuis, and dublin.** *Infect Immun* 1999, **67**:2611-2614.
  106. Chu C, Chiu CH: **Evolution of the virulence plasmids of non-typhoid *Salmonella* and its association with antimicrobial resistance.** *Microbes Infect* 2006, **8**:1931-1936.
  107. Rotger R, Casadesús J: **The virulence plasmids of *Salmonella*.** *Int Microbiol* 1999, **2**:177-184.
  108. Marcus SL, Brumell JH, Pfeifer CG, Finlay BB: ***Salmonella* pathogenicity islands: big virulence in small packages.** *Microbes Infect* 2000, **2**:145-156.
  109. Barth S, Bauerfeind R: **[Virulence plasmids of *Salmonella enterica*--incidence and properties].** *Berl Munch Tierarztl Wochenschr* 2005, **118**:8-23.

110. Wain J, Diem Nga LT, Kidgell C, James K, Fortune S, Song Diep T, Ali T, P OG, Parry C, Parkhill J, et al.: **Molecular analysis of *incHII* antimicrobial resistance plasmids from *Salmonella* serovar Typhi strains associated with typhoid fever.** *Antimicrob Agents Chemother* 2003, **47**:2732-2739.
111. Fica A, Fernandez-Beros ME, Aron-Hott L, Rivas A, D'Ottone K, Chumpitaz J, Guevara JM, Rodriguez M, Cabello F: **Antibiotic-resistant *Salmonella* typhi from two outbreaks: few ribotypes and IS200 types harbor Inc HI1 plasmids.** *Microb Drug Resist* 1997, **3**:339-343.
112. Fierer J, Guiney DG: **Diverse virulence traits underlying different clinical outcomes of *Salmonella* infection.** *J Clin Invest* 2001, **107**:775-780.
113. Fàbrega A, Vila J: ***Salmonella enterica* serovar Typhimurium skills to succeed in the host: virulence and regulation.** *Clin Microbiol Rev* 2013, **26**:308-341.
114. Hensel M, Shea JE, Bäumler AJ, Gleeson C, Blattner F, Holden DW: **Analysis of the boundaries of *Salmonella* pathogenicity island 2 and the corresponding chromosomal region of *Escherichia coli* K-12.** *J Bacteriol* 1997, **179**:1105-1111.
115. Blanc-Potard AB, Groisman EA: **The *Salmonella selC* locus contains a pathogenicity island mediating intramacrophage survival.** *Embo j* 1997, **16**:5376-5385.
116. Wong KK, McClelland M, Stillwell LC, Sisk EC, Thurston SJ, Saffer JD: **Identification and sequence analysis of a 27-kilobase chromosomal fragment containing a *Salmonella* pathogenicity island located at 92 minutes on the chromosome map of *Salmonella enterica* serovar typhimurium LT2.** *Infect Immun* 1998, **66**:3365-3371.
117. Shah DH, Zhou X, Kim HY, Call DR, Guard J: **Transposon mutagenesis of *Salmonella enterica* serovar Enteritidis identifies genes that contribute to invasiveness in human and chicken cells and survival in egg albumen.** *Infect Immun* 2012, **80**:4203-4215.
118. Folkesson A, Löfdahl S, Normark S: **The *Salmonella enterica* subspecies I specific centisome 7 genomic island encodes novel protein families present in bacteria living in close contact with eukaryotic cells.** *Res Microbiol* 2002, **153**:537-545.

119. Hansen-Wester I, Hensel M: **Genome-based identification of chromosomal regions specific for *Salmonella* spp.** *Infect Immun* 2002, **70**:2351-2360.
120. Pickard D, Wain J, Baker S, Line A, Chohan S, Fookes M, Barron A, Gaora PO, Chabalgoity JA, Thanky N, et al.: **Composition, acquisition, and distribution of the Vi exopolysaccharide-encoding *Salmonella enterica* pathogenicity island SPI-7.** *J Bacteriol* 2003, **185**:5055-5065.
121. Morris C, Tam CK, Wallis TS, Jones PW, Hackett J: ***Salmonella enterica* serovar Dublin strains which are Vi antigen-positive use type IVB pili for bacterial self-association and human intestinal cell entry.** *Microb Pathog* 2003, **35**:279-284.
122. Zhang XL, Morris C, Hackett J: **Molecular cloning, nucleotide sequence, and function of a site-specific recombinase encoded in the major 'pathogenicity island' of *Salmonella typhi*.** *Gene* 1997, **202**:139-146.
123. Parkhill J, Dougan G, James KD, Thomson NR, Pickard D, Wain J, Churcher C, Mungall KL, Bentley SD, Holden MT, et al.: **Complete genome sequence of a multiple drug resistant *Salmonella enterica* serovar Typhi CT18.** *Nature* 2001, **413**:848-852.
124. Velásquez JC, Hidalgo AA, Villagra N, Santiviago CA, Mora GC, Fuentes JA: **SPI-9 of *Salmonella enterica* serovar Typhi is constituted by an operon positively regulated by RpoS and contributes to adherence to epithelial cells in culture.** *Microbiology (Reading)* 2016, **162**:1367-1378.
125. Bishop AL, Baker S, Jenks S, Fookes M, Gaora PO, Pickard D, Anjum M, Farrar J, Hien TT, Ivens A, et al.: **Analysis of the hypervariable region of the *Salmonella enterica* genome associated with tRNA(LeuX).** *Journal of bacteriology* 2005, **187**:2469-2482.
126. Shah DH, Lee MJ, Park JH, Lee JH, Eo SK, Kwon JT, Chae JS: **Identification of *Salmonella gallinarum* virulence genes in a chicken infection model using PCR-based signature-tagged mutagenesis.** *Microbiology (Reading)* 2005, **151**:3957-3968.
127. Townsend SM, Kramer NE, Edwards R, Baker S, Hamlin N, Simmonds M, Stevens K, Maloy S, Parkhill J, Dougan G, et al.: ***Salmonella enterica* serovar Typhi possesses a unique repertoire of fimbrial gene sequences.** *Infect Immun* 2001, **69**:2894-2901.

128. Spanò S, Ugalde JE, Galán JE: **Delivery of a *Salmonella* Typhi exotoxin from a host intracellular compartment.** *Cell Host Microbe* 2008, **3**:30-38.
129. Tomljenovic-Berube AM, Henriksbo B, Porwollik S, Cooper CA, Tuinema BR, McClelland M, Coombes BK: **Mapping and regulation of genes within *Salmonella* pathogenicity island 12 that contribute to in vivo fitness of *Salmonella enterica* Serovar Typhimurium.** *Infect Immun* 2013, **81**:2394-2404.
130. Vernikos GS, Parkhill J: **Interpolated variable order motifs for identification of horizontally acquired DNA: revisiting the *Salmonella* pathogenicity islands.** *Bioinformatics* 2006, **22**:2196-2203.
131. Faucher SP, Forest C, Béland M, Daigle F: **A novel PhoP-regulated locus encoding the cytolysin ClyA and the secreted invasin TaiA of *Salmonella enterica* serovar Typhi is involved in virulence.** *Microbiology (Reading)* 2009, **155**:477-488.
132. Fuentes JA, Villagra N, Castillo-Ruiz M, Mora GC: **The *Salmonella* Typhi *hlyE* gene plays a role in invasion of cultured epithelial cells and its functional transfer to *S. Typhimurium* promotes deep organ infection in mice.** *Res Microbiol* 2008, **159**:279-287.
133. Urrutia IM, Fuentes JA, Valenzuela LM, Ortega AP, Hidalgo AA, Mora GC: ***Salmonella* Typhi *shdA*: pseudogene or allelic variant?** *Infect Genet Evol* 2014, **26**:146-152.
134. Kingsley RA, Humphries AD, Weening EH, De Zoete MR, Winter S, Papaconstantinopoulou A, Dougan G, Bäumlér AJ: **Molecular and phenotypic analysis of the CS54 island of *Salmonella enterica* serotype typhimurium: identification of intestinal colonization and persistence determinants.** *Infect Immun* 2003, **71**:629-640.
135. Kropinski AM, Sulakvelidze A, Konczyk P, Poppe C: ***Salmonella* phages and prophages--genomics and practical aspects.** *Methods Mol Biol* 2007, **394**:133-175.
136. Boyd EF, Brüssow H: **Common themes among bacteriophage-encoded virulence factors and diversity among the bacteriophages involved.** *Trends Microbiol* 2002, **10**:521-529.

137. Rudolph MG, Weise C, Mirol S, Hillenbrand B, Bader B, Wittinghofer A, Hardt WD: **Biochemical analysis of SopE from *Salmonella typhimurium*, a highly efficient guanosine nucleotide exchange factor for RhoGTPases.** *J Biol Chem* 1999, **274**:30501-30509.
138. Hardt WD, Chen LM, Schuebel KE, Bustelo XR, Galán JE: ***S. typhimurium* encodes an activator of Rho GTPases that induces membrane ruffling and nuclear responses in host cells.** *Cell* 1998, **93**:815-826.
139. Mirol S, Ehrbar K, Weissmüller A, Prager R, Tschäpe H, Rüssmann H, Hardt WD: ***Salmonella* host cell invasion emerged by acquisition of a mosaic of separate genetic elements, including *Salmonella* pathogenicity island 1 (SPI1), SPI5, and sopE2.** *J Bacteriol* 2001, **183**:2348-2358.
140. Mirol S, Rabsch W, Rohde M, Stender S, Tschäpe H, Rüssmann H, Igwe E, Hardt WD: **Isolation of a temperate bacteriophage encoding the type III effector protein SopE from an epidemic *Salmonella typhimurium* strain.** *Proc Natl Acad Sci U S A* 1999, **96**:9845-9850.
141. Figueroa-Bossi N, Uzzau S, Maloriol D, Bossi L: **Variable assortment of prophages provides a transferable repertoire of pathogenic determinants in *Salmonella*.** *Mol Microbiol* 2001, **39**:260-271.
142. Figueroa-Bossi N, Bossi L: **Inducible prophages contribute to *Salmonella* virulence in mice.** *Mol Microbiol* 1999, **33**:167-176.
143. Uzzau S, Figueroa-Bossi N, Rubino S, Bossi L: **Epitope tagging of chromosomal genes in *Salmonella*.** *Proc Natl Acad Sci U S A* 2001, **98**:15264-15269.
144. Stanley TL, Ellermeier CD, Slauch JM: **Tissue-specific gene expression identifies a gene in the lysogenic phage Gifsy-1 that affects *Salmonella enterica* serovar typhimurium survival in Peyer's patches.** *J Bacteriol* 2000, **182**:4406-4413.
145. Worley MJ, Ching KH, Heffron F: ***Salmonella* SsrB activates a global regulon of horizontally acquired genes.** *Mol Microbiol* 2000, **36**:749-761.
146. Miao EA, Miller SI: **A conserved amino acid sequence directing intracellular type III secretion by *Salmonella typhimurium*.** *Proc Natl Acad Sci U S A* 2000, **97**:7539-7544.



147. Ho TD, Figueroa-Bossi N, Wang M, Uzzau S, Bossi L, Slauch JM: **Identification of GtgE, a novel virulence factor encoded on the Gifsy-2 bacteriophage of *Salmonella enterica* serovar Typhimurium.** *J Bacteriol* 2002, **184**:5234-5239.
148. De Groote MA, Ochsner UA, Shiloh MU, Nathan C, McCord JM, Dinauer MC, Libby SJ, Vazquez-Torres A, Xu Y, Fang FC: **Periplasmic superoxide dismutase protects *Salmonella* from products of phagocyte NADPH-oxidase and nitric oxide synthase.** *Proc Natl Acad Sci U S A* 1997, **94**:13997-14001.
149. Farrant JL, Sansone A, Canvin JR, Pallen MJ, Langford PR, Wallis TS, Dougan G, Kroll JS: **Bacterial copper- and zinc-cofactored superoxide dismutase contributes to the pathogenesis of systemic salmonellosis.** *Mol Microbiol* 1997, **25**:785-796.
150. Miao EA, Scherer CA, Tsolis RM, Kingsley RA, Adams LG, Bäumlér AJ, Miller SI: ***Salmonella typhimurium* leucine-rich repeat proteins are targeted to the SPI1 and SPI2 type III secretion systems.** *Mol Microbiol* 1999, **34**:850-864.
151. Tsolis RM, Townsend SM, Miao EA, Miller SI, Ficht TA, Adams LG, Bäumlér AJ: **Identification of a putative *Salmonella enterica* serotype typhimurium host range factor with homology to IpaH and YopM by signature-tagged mutagenesis.** *Infect Immun* 1999, **67**:6385-6393.
152. Haraga A, Miller SI: **A *Salmonella enterica* serovar typhimurium translocated leucine-rich repeat effector protein inhibits NF-kappa B-dependent gene expression.** *Infect Immun* 2003, **71**:4052-4058.
153. Stender S, Friebel A, Linder S, Rohde M, Mirol S, Hardt WD: **Identification of SopE2 from *Salmonella typhimurium*, a conserved guanine nucleotide exchange factor for Cdc42 of the host cell.** *Mol Microbiol* 2000, **36**:1206-1221.
154. Miao EA, Brittnacher M, Haraga A, Jeng RL, Welch MD, Miller SI: ***Salmonella* effectors translocated across the vacuolar membrane interact with the actin cytoskeleton.** *Mol Microbiol* 2003, **48**:401-415.
155. Boyd EF, Porwollik S, Blackmer F, McClelland M: **Differences in gene content among *Salmonella enterica* serovar typhi isolates.** *J Clin Microbiol* 2003, **41**:3823-3828.
156. Porwollik S, McClelland M: **Lateral gene transfer in *Salmonella*.** *Microbes Infect* 2003, **5**:977-989.

157. Agron PG, Walker RL, Kinde H, Sawyer SJ, Hayes DC, Wollard J, Andersen GL:  
**Identification by subtractive hybridization of sequences specific for *Salmonella enterica* serovar enteritidis.** *Applied and environmental microbiology* 2001, **67**:4984-4991.
158. Santiviago CA, Blondel CJ, Quezada CP, Silva CA, Tobar PM, Porwollik S, McClelland M, Andrews-Polymeris HL, Toro CS, Zaldívar M, et al.:  
**Spontaneous excision of the *Salmonella enterica* serovar Enteritidis-specific defective prophage-like element phiSE14.** *Journal of bacteriology* 2010, **192**:2246-2254.
159. Thomson NR, Clayton DJ, Windhorst D, Vernikos G, Davidson S, Churcher C, Quail MA, Stevens M, Jones MA, Watson M, et al.: **Comparative genome analysis of *Salmonella* Enteritidis PT4 and *Salmonella* Gallinarum 287/91 provides insights into evolutionary and host adaptation pathways.** *Genome Res* 2008, **18**:1624-1637.
160. Lee K-M, Runyon M, Herrman TJ, Phillips R, Hsieh J: **Review of *Salmonella* detection and identification methods: Aspects of rapid emergency response and food safety.** *J Food control* 2015, **47**:264-276.
161. Carrique-Mas JJ, Davies RH: **Sampling and bacteriological detection of *Salmonella* in poultry and poultry premises: a review.** *Rev Sci Tech* 2008, **27**:665-677.
162. Ruiz J, Núñez ML, Díaz J, Sempere MA, Gómez J, Usera MA: **Note: comparison of media for the isolation of lactose-positive *Salmonella*.** *J Appl Bacteriol* 1996, **81**:571-574.
163. Mallinson ET, Miller RG, de Rezende CE, Ferris KE, deGraft-Hanson J, Joseph SW: **Improved plating media for the detection of *Salmonella* species with typical and atypical hydrogen sulfide production.** *J Vet Diagn Invest* 2000, **12**:83-87.
164. Arrach N, Porwollik S, Cheng P, Cho A, Long F, Choi SH, McClelland M:  
***Salmonella* serovar identification using PCR-based detection of gene presence and absence.** *J Clin Microbiol* 2008, **46**:2581-2589.
165. Hoorfar J, Baggesen DL, Porting PH: **A PCR-based strategy for simple and rapid identification of rough presumptive *Salmonella* isolates.** *J Microbiol Methods* 1999, **35**:77-84.

166. Schrader KN, Fernandez-Castro A, Cheung WK, Crandall CM, Abbott SL: **Evaluation of commercial antisera for *Salmonella* serotyping.** *J Clin Microbiol* 2008, **46**:685-688.
167. Herrera-León S, Ramiro R, Arroyo M, Díez R, Usera MA, Echeita MA: **Blind comparison of traditional serotyping with three multiplex PCRs for the identification of *Salmonella* serotypes.** *Res Microbiol* 2007, **158**:122-127.
168. McQuiston JR, Waters RJ, Dinsmore BA, Mikoleit ML, Fields PI: **Molecular determination of H antigens of *Salmonella* by use of a microsphere-based liquid array.** *J Clin Microbiol* 2011, **49**:565-573.
169. Herrera-León S, McQuiston JR, Usera MA, Fields PI, Garaizar J, Echeita MA: **Multiplex PCR for distinguishing the most common phase-1 flagellar antigens of *Salmonella* spp.** *J Clin Microbiol* 2004, **42**:2581-2586.
170. Echeita MA, Herrera S, Garaizar J, Usera MA: **Multiplex PCR-based detection and identification of the most common *Salmonella* second-phase flagellar antigens.** *Res Microbiol* 2002, **153**:107-113.
171. Braun SD, Ziegler A, Methner U, Slickers P, Keiling S, Monecke S, Ehricht R: **Fast DNA serotyping and antimicrobial resistance gene determination of *salmonella enterica* with an oligonucleotide microarray-based assay.** *PLoS One* 2012, **7**:e46489.
172. Yoshida C, Franklin K, Konczy P, McQuiston JR, Fields PI, Nash JH, Taboada EN, Rahn K: **Methodologies towards the development of an oligonucleotide microarray for determination of *Salmonella* serotypes.** *J Microbiol Methods* 2007, **70**:261-271.
173. Ranieri ML, Shi C, Moreno Switt AI, den Bakker HC, Wiedmann M: **Comparison of typing methods with a new procedure based on sequence characterization for *Salmonella* serovar prediction.** *J Clin Microbiol* 2013, **51**:1786-1797.
174. Franklin K, Lingohr EJ, Yoshida C, Anjum M, Bodrossy L, Clark CG, Kropinski AM, Karmali MA: **Rapid genoserotyping tool for classification of *Salmonella* serovars.** *J Clin Microbiol* 2011, **49**:2954-2965.
175. Kim S, Frye JG, Hu J, Fedorka-Cray PJ, Gautom R, Boyle DS: **Multiplex PCR-based method for identification of common clinical serotypes of *Salmonella enterica* subsp. *enterica*.** *Journal of clinical microbiology* 2006, **44**:3608-3615.

176. Peterson G, Gerdes B, Berges J, Nagaraja TG, Frye JG, Boyle DS, Narayanan S: **Development of microarray and multiplex polymerase chain reaction assays for identification of serovars and virulence genes in *Salmonella enterica* of human or animal origin.** *J Vet Diagn Invest* 2010, **22**:559-569.
177. Laing CR, Whiteside MD, Gannon VPJ: **Pan-genome Analyses of the Species *Salmonella enterica*, and Identification of Genomic Markers Predictive for Species, Subspecies, and Serovar.** *Frontiers in microbiology* 2017, **8**:1345.
178. Zou QH, Li RQ, Liu GR, Liu SL: **Genotyping of *Salmonella* with lineage-specific genes: correlation with serotyping.** *Int J Infect Dis* 2016, **49**:134-140.
179. Moore MM, Feist MD: **Real-time PCR method for *Salmonella* spp. targeting the *stn* gene.** *J Appl Microbiol* 2007, **102**:516-530.
180. Thompson CP, Doak AN, Amirani N, Schroeder EA, Wright J, Kariyawasam S, Lamendella R, Shariat NW: **High-Resolution Identification of Multiple *Salmonella* Serovars in a Single Sample by Using CRISPR-SeroSeq.** *Appl Environ Microbiol* 2018, **84**.
181. Farrell JJ, Doyle LJ, Addison RM, Reller LB, Hall GS, Procop GW: **Broad-range (pan) *Salmonella* and *Salmonella* serotype typhi-specific real-time PCR assays: potential tools for the clinical microbiologist.** *Am J Clin Pathol* 2005, **123**:339-345.
182. Porwollik S, Boyd EF, Choy C, Cheng P, Florea L, Proctor E, McClelland M: **Characterization of *Salmonella enterica* subspecies I genovars by use of microarrays.** *J Bacteriol* 2004, **186**:5883-5898.
183. Porwollik S, Santiviago CA, Cheng P, Florea L, Jackson S, McClelland M: **Differences in gene content between *Salmonella enterica* serovar Enteritidis isolates and comparison to closely related serovars Gallinarum and Dublin.** *J Bacteriol* 2005, **187**:6545-6555.
184. Malorny B, Hoorfar J, Bunge C, Helmuth R: **Multicenter validation of the analytical accuracy of *Salmonella* PCR: towards an international standard.** *Appl Environ Microbiol* 2003, **69**:290-296.
185. Chiu CH, Ou JT: **Rapid identification of *Salmonella* serovars in feces by specific detection of virulence genes, *invA* and *spvC*, by an enrichment broth culture-multiplex PCR combination assay.** *J Clin Microbiol* 1996, **34**:2619-2622.

186. Rajtak U, Leonard N, Bolton D, Fanning S: **A real-time multiplex SYBR Green I polymerase chain reaction assay for rapid screening of *salmonella* serotypes prevalent in the European Union.** *Foodborne Pathog Dis* 2011, **8**:769-780.
187. O'Regan E, McCabe E, Burgess C, McGuinness S, Barry T, Duffy G, Whyte P, Fanning S: **Development of a real-time multiplex PCR assay for the detection of multiple *Salmonella* serotypes in chicken samples.** *BMC Microbiol* 2008, **8**:156.
188. Maiden MC, Bygraves JA, Feil E, Morelli G, Russell JE, Urwin R, Zhang Q, Zhou J, Zurth K, Caugant DA, et al.: **Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms.** *Proc Natl Acad Sci U S A* 1998, **95**:3140-3145.
189. Kotetishvili M, Stine OC, Kreger A, Morris JG, Jr., Sulakvelidze A: **Multilocus sequence typing for characterization of clinical and environmental *Salmonella* strains.** *J Clin Microbiol* 2002, **40**:1626-1635.
190. Jolley KA, Bliss CM, Bennett JS, Bratcher HB, Brehony C, Colles FM, Wimalarathna H, Harrison OB, Sheppard SK, Cody AJ, et al.: **Ribosomal multilocus sequence typing: universal characterization of bacteria from domain to strain.** *Microbiology (Reading)* 2012, **158**:1005-1015.
191. Sheppard SK, Jolley KA, Maiden MC: **A Gene-By-Gene Approach to Bacterial Population Genomics: Whole Genome MLST of *Campylobacter*.** *Genes (Basel)* 2012, **3**:261-277.
192. Maiden MC, Jansen van Rensburg MJ, Bray JE, Earle SG, Ford SA, Jolley KA, McCarthy ND: **MLST revisited: the gene-by-gene approach to bacterial genomics.** *Nat Rev Microbiol* 2013, **11**:728-736.
193. Yoshida CE, Kruczkiewicz P, Laing CR, Lingohr EJ, Gannon VP, Nash JH, Taboada EN: **The *Salmonella In Silico* Typing Resource (SISTR): An Open Web-Accessible Tool for Rapidly Typing and Subtyping Draft *Salmonella* Genome Assemblies.** *PLoS One* 2016, **11**:e0147101.
194. Zhang S, den Bakker HC, Li S, Chen J, Dinsmore BA, Lane C, Lauer AC, Fields PI, Deng X: **SeqSero2: Rapid and Improved *Salmonella* Serotype Determination Using Whole-Genome Sequencing Data.** *Appl Environ Microbiol* 2019, **85**.

195. Zhang S, Yin Y, Jones MB, Zhang Z, Deatherage Kaiser BL, Dinsmore BA, Fitzgerald C, Fields PI, Deng X: ***Salmonella* serotype determination utilizing high-throughput genome sequencing data.** *Journal of clinical microbiology* 2015, **53**:1685-1692.
196. Diep B, Barretto C, Portmann AC, Fournier C, Karczmarek A, Voets G, Li S, Deng X, Klijn A: ***Salmonella* Serotyping; Comparison of the Traditional Method to a Microarray-Based Method and an *in silico* Platform Using Whole Genome Sequencing Data.** *Front Microbiol* 2019, **10**:2554.
197. Wang Y, Wang Y, Zhang L, Liu D, Luo L, Li H, Cao X, Liu K, Xu J, Ye C: **Multiplex, Rapid, and Sensitive Isothermal Detection of Nucleic-Acid Sequence by Endonuclease Restriction-Mediated Real-Time Multiple Cross Displacement Amplification.** *Frontiers in microbiology* 2016, **7**:753.
198. Wang Y, Wang Y, Ma AJ, Li DX, Luo LJ, Liu DX, Jin D, Liu K, Ye CY: **Rapid and Sensitive Isothermal Detection of Nucleic-acid Sequence by Multiple Cross Displacement Amplification.** *Sci Rep* 2015, **5**:11902.
199. Law JW, Ab Mutalib NS, Chan KG, Lee LH: **Rapid methods for the detection of foodborne bacterial pathogens: principles, applications, advantages and limitations.** *Front Microbiol* 2014, **5**:770.
200. Azinheiro S, Carvalho J, Prado M, Garrido-Maestu A: **Evaluation of Different Genetic Targets for *Salmonella enterica* Serovar Enteritidis and Typhimurium, Using Loop-Mediated Isothermal AMPLification for Detection in Food Samples.** *Frontiers in Sustainable Food Systems* 2018, **2**:5.
201. Garrido-Maestu A, Fuciños P, Azinheiro S, Carvalho J, Prado M: **Systematic loop-mediated isothermal amplification assays for rapid detection and characterization of *Salmonella* spp., Enteritidis and Typhimurium in food samples.** *Food Control* 2017, **80**:297-306.
202. Gong J, Zhuang L, Zhu C, Shi S, Zhang D, Zhang L, Yu Y, Dou X, Xu B, Wang C: **Loop-Mediated Isothermal Amplification of the *sefA* Gene for Rapid Detection of *Salmonella* Enteritidis and *Salmonella* Gallinarum in Chickens.** *Foodborne Pathog Dis* 2016, **13**:177-181.
203. Hu L, Ma LM, Zheng S, He X, Hammack TS, Brown EW, Zhang G: **Development of a novel loop-mediated isothermal amplification (LAMP) assay for the**

- detection of *Salmonella* ser. Enteritidis from egg products. *Food control* 2018, **88**:190-197.
204. Okamura M, Ohba Y, Kikuchi S, Suzuki A, Tachizaki H, Takehara K, Ikedo M, Kojima T, Nakamura M: **Loop-mediated isothermal amplification for the rapid, sensitive, and specific detection of the O9 group of *Salmonella* in chickens.** *Vet Microbiol* 2008, **132**:197-204.
205. Okamura M, Ohba Y, Kikuchi S, Takehara K, Ikedo M, Kojima T, Nakamura M: **Rapid, sensitive, and specific detection of the O4 group of *Salmonella enterica* by loop-mediated isothermal amplification.** *Avian Dis* 2009, **53**:216-221.
206. Pavan Kumar P, Agarwal R, Thomas P, Sailo B, Prasannavadhana A, Kumar A, Kataria J, Singh D: **Rapid detection of *Salmonella enterica* subspecies *enterica* serovar Typhimurium by loop mediated isothermal amplification (LAMP) test from field chicken meat samples.** *Food biotechnology* 2014, **28**:50-62.
207. Ravan H, Yazdanparast R: **Development of a new loop-mediated isothermal amplification assay for prt (*rfbS*) gene to improve the identification of *Salmonella* serogroup D.** *World J Microbiol Biotechnol* 2012, **28**:2101-2106.
208. Yang JL, Ma GP, Yang R, Yang SQ, Fu LZ, Cheng AC, Wang MS, Zhang SH, Shen KF, Jia RY, et al.: **Simple and rapid detection of *Salmonella* serovar Enteritidis under field conditions by loop-mediated isothermal amplification.** *Journal of applied microbiology* 2010, **109**:1715-1723.
209. Baker S, The HC: **Recent insights into *Shigella*.** *Curr Opin Infect Dis* 2018, **31**:449-454.
210. Kotloff KL, Riddle MS, Platts-Mills JA, Pavlinac P, Zaidi AKM: **Shigellosis.** *Lancet* 2018, **391**:801-812.
211. Malani PN: **Mandell, Douglas, and Bennett's principles and practice of infectious diseases.** *JAMA* 2010, **304**:2067-2071.
212. Kaper JB, Nataro JP, Mobley HL: **Pathogenic *Escherichia coli*.** *Nat Rev Microbiol* 2004, **2**:123-140.
213. Nataro JP, Kaper JB: **Diarrheagenic *escherichia coli*.** *Clinical microbiology reviews* 1998, **11**:142-201.

214. van den Beld MJ, Reubsaet FA: **Differentiation between *Shigella*, enteroinvasive *Escherichia coli* (EIEC) and noninvasive *Escherichia coli*.** *Eur J Clin Microbiol Infect Dis* 2012, **31**:899-904.
215. Yabuuchi E: **Bacillus dysentericus (sic) 1897 was the first taxonomic rather than Bacillus dysenteriae 1898.** *Int J Syst Evol Microbiol* 2002, **52**:1041.
216. Sahl JW, Morris CR, Emberger J, Fraser CM, Ochieng JB, Juma J, Fields B, Breiman RF, Gilmour M, Nataro JP, et al.: **Defining the phylogenomics of *Shigella* species: a pathway to diagnostics.** *Journal of clinical microbiology* 2015, **53**:951-960.
217. Lindberg AA, Kärnell A, Weintraub A: **The lipopolysaccharide of *Shigella* bacteria as a virulence factor.** *Rev Infect Dis* 1991, **13 Suppl 4**:S279-284.
218. Liu B, Knirel YA, Feng L, Perepelov AV, Senchenkova SN, Wang Q, Reeves PR, Wang L: **Structure and genetics of *Shigella* O antigens.** *FEMS microbiology reviews* 2008, **32**:627-653.
219. Teh MY, Furevi A, Widmalm G, Morona R: **Influence of *Shigella flexneri* 2a O Antigen Acetylation on Its Bacteriophage Sf6 Receptor Activity and Bacterial Interaction with Human Cells.** *J Bacteriol* 2020, **202**.
220. Knirel YA, Sun Q, Senchenkova SN, Perepelov AV, Shashkov AS, Xu J: **O-antigen modifications providing antigenic diversity of *Shigella flexneri* and underlying genetic mechanisms.** *Biochemistry (Mosc)* 2015, **80**:901-914.
221. Carlin NI, Rahman M, Sack DA, Zaman A, Kay B, Lindberg AA: **Use of monoclonal antibodies to type *Shigella flexneri* in Bangladesh.** *J Clin Microbiol* 1989, **27**:1163-1166.
222. Ochman H, Whittam TS, Causant DA, Selander RK: **Enzyme polymorphism and genetic population structure in *Escherichia coli* and *Shigella*.** *J Gen Microbiol* 1983, **129**:2715-2726.
223. Pupo GM, Karaolis DK, Lan R, Reeves PR: **Evolutionary relationships among pathogenic and nonpathogenic *Escherichia coli* strains inferred from multilocus enzyme electrophoresis and mdh sequence studies.** *Infect Immun* 1997, **65**:2685-2692.
224. Hartl DL, Dykhuizen DE: **The population genetics of *Escherichia coli*.** *J Annual review of genetics* 1984, **18**:31-68.
225. Ewing WH: **SHIGELLA NOMENCLATURE.** *J Bacteriol* 1949, **57**:633-638.



226. DuPont HL, Formal SB, Hornick RB, Snyder MJ, Libonati JP, Sheahan DG, LaBrec EH, Kalas JP: **Pathogenesis of *Escherichia coli* diarrhea.** *N Engl J Med* 1971, **285**:1-9.
227. Kauffmann F: **The serology of the *coli* group.** *J Immunol* 1947, **57**:71-100.
228. Valvano MA: **Export of O-specific lipopolysaccharide.** *Front Biosci* 2003, **8**:s452-471.
229. Liu B, Furevi A, Perepelov AV, Guo X, Cao H, Wang Q, Reeves PR, Knirel YA, Wang L, Widmalm G: **Structure and genetics of *Escherichia coli* O antigens.** *FEMS microbiology reviews* 2020, **44**:655-683.
230. Iguchi A, Iyoda S, Kikuchi T, Ogura Y, Katsura K, Ohnishi M, Hayashi T, Thomson NR: **A complete view of the genetic diversity of the *Escherichia coli* O-antigen biosynthesis gene cluster.** *DNA Res* 2015, **22**:101-107.
231. Samuel G, Reeves P: **Biosynthesis of O-antigens: genes and pathways involved in nucleotide sugar precursor synthesis and O-antigen assembly.** *Carbohydr Res* 2003, **338**:2503-2519.
232. Feng L, Liu B, Liu Y, Ratiner YA, Hu B, Li D, Zong X, Xiong W, Wang L: **A genomic islet mediates flagellar phase variation in *Escherichia coli* strains carrying the flagellin-specifying locus *flk*.** *J Bacteriol* 2008, **190**:4470-4477.
233. Ingle DJ, Valcanis M, Kuzevski A, Tauschek M, Inouye M, Stinear T, Levine MM, Robins-Browne RM, Holt KE: ***In silico* serotyping of *E. coli* from short read data identifies limited novel O-loci but extensive diversity of O:H serotype combinations within and between pathogenic lineages.** *Microb Genom* 2016, **2**:e000064.
234. Ratiner YA: **New flagellin-specifying genes in some *Escherichia coli* strains.** *J Bacteriol* 1998, **180**:979-984.
235. Ratiner YA, Sihvonen LM, Liu Y, Wang L, Siitonen A: **Alteration of flagellar phenotype of *Escherichia coli* strain P12b, the standard type strain for flagellar antigen H17, possessing a new non-*fliC* flagellin gene *flnA*, and possible loss of original flagellar phenotype and genotype in the course of subculturing through semisolid media.** *Arch Microbiol* 2010, **192**:267-278.
236. Tominaga A: **Characterization of six flagellin genes in the H3, H53 and H54 standard strains of *Escherichia coli*.** *Genes Genet Syst* 2004, **79**:1-8.

237. Tominaga A, Kutsukake K: **Expressed and cryptic flagellin genes in the H44 and H55 type strains of *Escherichia coli*.** *Genes Genet Syst* 2007, **82**:1-8.
238. Wang L, Rothmund D, Curd H, Reeves PR: **Species-wide variation in the *Escherichia coli* flagellin (H-antigen) gene.** *Journal of bacteriology* 2003, **185**:2936-2943.
239. Ørskov F, Ørskov I: **2 Serotyping of *Escherichia coli*.** *Methods in microbiology* 1984, **14**:43-112.
240. ørskov I, ørskov F, Bettelheim KA, Chandler ME: **Two new *Escherichia coli* o antigens, o162 and o163, and one new h antigen, h56. withdrawal of h antigen h50.** *Acta Pathol Microbiol Scand B* 1975, **83**:121-124.
241. Scheutz F, Cheasty T, Woodward D, Smith HR: **Designation of O174 and O175 to temporary O groups OX3 and OX7, and six new *E. coli* O groups that include Verocytotoxin-producing *E. coli* (VTEC): O176, O177, O178, O179, O180 and O181.** *Apmis* 2004, **112**:569-584.
242. Yang S, Xi D, Jing F, Kong D, Wu J, Feng L, Cao B, Wang L: **Genetic diversity of K-antigen gene clusters of *Escherichia coli* and their molecular typing using a suspension array.** *Can J Microbiol* 2018, **64**:231-241.
243. Whitfield C, Roberts IS: **Structure, assembly and regulation of expression of capsules in *Escherichia coli*.** *Mol Microbiol* 1999, **31**:1307-1319.
244. Herzig CTA, Fleischauer AT, Lackey B, Lee N, Lawson T, Moore ZS, Hergert J, Mobley V, MacFarquhar J, Morrison T, et al.: **Notes from the Field: Enteroinvasive *Escherichia coli* Outbreak Associated with a Potluck Party - North Carolina, June-July 2018.** *MMWR Morbidity and mortality weekly report* 2019, **68**:183-184.
245. Pasqua M, Michelacci V, Di Martino ML, Tozzoli R, Grossi M, Colonna B, Morabito S, Prosseda G: **The Intriguing Evolutionary Journey of Enteroinvasive *E. coli* (EIEC) toward Pathogenicity.** *Frontiers in microbiology* 2017, **8**:2390.
246. Croxen MA, Law RJ, Scholz R, Keeney KM, Wlodarska M, Finlay BB: **Recent advances in understanding enteric pathogenic *Escherichia coli*.** *Clin Microbiol Rev* 2013, **26**:822-880.
247. Kotloff KL, Blackwelder WC, Nasrin D, Nataro JP, Farag TH, van Eijk A, Adegbola RA, Alonso PL, Breiman RF, Faruque AS, et al.: **The Global Enteric**

- Multicenter Study (GEMS) of diarrheal disease in infants and young children in developing countries: epidemiologic and clinical methods of the case/control study.** *Clin Infect Dis* 2012, **55** Suppl 4:S232-245.
248. Kotloff KL, Nataro JP, Blackwelder WC, Nasrin D, Farag TH, Panchalingam S, Wu Y, Sow SO, Sur D, Breiman RF, et al.: **Burden and aetiology of diarrhoeal disease in infants and young children in developing countries (the Global Enteric Multicenter Study, GEMS): a prospective, case-control study.** *Lancet* 2013, **382**:209-222.
249. Hawkey J, Paranagama K, Baker KS, Bengtsson RJ, Weill FX, Thomson NR, Baker S, Cerdeira L, Iqbal Z, Hunt M, et al.: **Global population structure and genotyping framework for genomic surveillance of the major dysentery pathogen, *Shigella sonnei*.** *Nat Commun* 2021, **12**:2684.
250. Anderson M, Sansonetti PJ, Marteyn BS: ***Shigella* Diversity and Changing Landscape: Insights for the Twenty-First Century.** *Front Cell Infect Microbiol* 2016, **6**:45.
251. Livio S, Strockbine NA, Panchalingam S, Tennant SM, Barry EM, Marohn ME, Antonio M, Hossain A, Mandomando I, Ochieng JB, et al.: ***Shigella* isolates from the global enteric multicenter study inform vaccine development.** *Clinical infectious diseases* 2014, **59**:933-941.
252. Ye C, Lan R, Xia S, Zhang J, Sun Q, Zhang S, Jing H, Wang L, Li Z, Zhou Z, et al.: **Emergence of a new multidrug-resistant serotype X variant in an epidemic clone of *Shigella flexneri*.** *Journal of clinical microbiology* 2010, **48**:419-426.
253. Levine MM, Kotloff KL, Barry EM, Pasetti MF, Sztein MB: **Clinical trials of *Shigella* vaccines: two steps forward and one step back on a long, hard road.** *Nat Rev Microbiol* 2007, **5**:540-553.
254. von Seidlein L, Kim DR, Ali M, Lee H, Wang X, Thiem VD, Canh DG, Chaicumpa W, Agtini MD, Hossain A, et al.: **A multicentre study of *Shigella* diarrhoea in six Asian countries: disease burden, clinical manifestations, and microbiology.** *PLoS Med* 2006, **3**:e353.
255. Collaborators. GDD: **Estimates of global, regional, and national morbidity, mortality, and aetiologies of diarrhoeal diseases: a systematic analysis for**

- the Global Burden of Disease Study 2015. *Lancet Infect Dis* 2017, **17**:909-948.
256. Baker KS, Dallman TJ, Ashton PM, Day M, Hughes G, Crook PD, Gilbert VL, Zittermann S, Allen VG, Howden BP, et al.: **Intercontinental dissemination of azithromycin-resistant shigellosis through sexual transmission: a cross-sectional study.** *Lancet Infect Dis* 2015, **15**:913-921.
257. Ingle DJ, Easton M, Valcanis M, Seemann T, Kwong JC, Stephens N, Carter GP, Gonçalves da Silva A, Adamopoulos J, Baines SL, et al.: **Co-circulation of Multidrug-resistant *Shigella* Among Men Who Have Sex With Men in Australia.** *Clin Infect Dis* 2019, **69**:1535-1544.
258. Kennedy FM, Astbury J, Needham JR, Cheasty T: **Shigellosis due to occupational contact with non-human primates.** *Epidemiol Infect* 1993, **110**:247-251.
259. Nizeyi JB, Innocent RB, Erume J, Kalema GR, Cranfield MR, Graczyk TK: **Campylobacteriosis, salmonellosis, and shigellosis in free-ranging human-habituated mountain gorillas of Uganda.** *J Wildl Dis* 2001, **37**:239-244.
260. Julian TR: **Environmental transmission of diarrheal pathogens in low and middle income countries.** *Environ Sci Process Impacts* 2016, **18**:944-955.
261. Aragón TJ, Vugia DJ, Shallow S, Samuel MC, Reingold A, Angulo FJ, Bradford WZ: **Case-control study of shigellosis in San Francisco: the role of sexual transmission and HIV infection.** *Clin Infect Dis* 2007, **44**:327-334.
262. Gilbert VL, Simms I, Jenkins C, Furegato M, Gobin M, Oliver I, Hart G, Gill ON, Hughes G: **Sex, drugs and smart phone applications: findings from semistructured interviews with men who have sex with men diagnosed with *Shigella flexneri* 3a in England and Wales.** *Sex Transm Infect* 2015, **91**:598-602.
263. Gomes TA, Elias WP, Scaletsky IC, Guth BE, Rodrigues JF, Piazza RM, Ferreira LC, Martinez MB: **Diarrheagenic *Escherichia coli*.** *Brazilian journal of microbiology* 2016, **47 Suppl 1**:3-30.
264. Vieira N, Bates SJ, Solberg OD, Ponce K, Howsmon R, Cevallos W, Trueba G, Riley L, Eisenberg JN: **High prevalence of enteroinvasive *Escherichia coli* isolated in a remote region of northern coastal Ecuador.** *Am J Trop Med Hyg* 2007, **76**:528-533.

265. Ud-Din A, Wahid S: **Relationship among *Shigella* spp. and enteroinvasive *Escherichia coli* (EIEC) and their differentiation.** *Brazilian journal of microbiology* 2014, **45**:1131-1138.
266. Escher M, Scavia G, Morabito S, Tozzoli R, Maugliani A, Cantoni S, Fracchia S, Bettati A, Casa R, Gesu GP, et al.: **A severe foodborne outbreak of diarrhoea linked to a canteen in Italy caused by enteroinvasive *Escherichia coli*, an uncommon agent.** *Epidemiology and infection* 2014, **142**:2559-2566.
267. Newitt S, MacGregor V, Robbins V, Bayliss L, Chattaway MA, Dallman T, Ready D, Aird H, Puleston R, Hawker J: **Two Linked Enteroinvasive *Escherichia coli* Outbreaks, Nottingham, UK, June 2014.** *Emerging infectious diseases* 2016, **22**:1178-1184.
268. Michelacci V, Prosseda G, Maugliani A, Tozzoli R, Sanchez S, Herrera-León S, Dallman T, Jenkins C, Caprioli A, Morabito S: **Characterization of an emergent clone of enteroinvasive *Escherichia coli* circulating in Europe.** *Clinical Microbiology* 2016, **22**:287. e211-287. e219.
269. Svenungsson B, Lagergren A, Ekwall E, Evengård B, Hedlund KO, Kärmell A, Löfdahl S, Svensson L, Weintraub A: **Enteropathogens in adult patients with diarrhea and healthy control subjects: a 1-year prospective study in a Swedish clinic for infectious diseases.** *Clin Infect Dis* 2000, **30**:770-778.
270. van den Beld MJC, Warmelink E, Friedrich AW, Reubsæet FAG, Schipper M, de Boer RF, Notermans DW, Petrignani MWF, van Zanten E, Rossen JWA, et al.: **Incidence, clinical implications and impact on public health of infections with *Shigella* spp. and entero-invasive *Escherichia coli* (EIEC): results of a multicenter cross-sectional study in the Netherlands during 2016-2017.** *BMC infectious diseases* 2019, **19**:1037.
271. Falkow S: **Activity of lysine decarboxylase as an aid in the identification of *Salmonellae* and *Shigellae*.** *Am J Clin Pathol* 1958, **29**:598-600.
272. Lan R, Alles MC, Donohoe K, Martinez MB, Reeves PR: **Molecular evolutionary relationships of enteroinvasive *Escherichia coli* and *Shigella* spp.** *Infection and immunity* 2004, **72**:5080-5088.
273. Goodman RE, Pickett MJ: **Delayed lactose fermentation by enterobacteriaceae.** *J Bacteriol* 1966, **92**:318-327.

274. Lan R, Lumb B, Ryan D, Reeves PR: **Molecular evolution of large virulence plasmid in *Shigella* clones and enteroinvasive *Escherichia coli*.** *Infection and immunity* 2001, **69**:6303-6309.
275. Buchrieser C, Glaser P, Rusniok C, Nedjari H, D'Hauteville H, Kunst F, Sansonetti P, Parsot C: **The virulence plasmid pWR100 and the repertoire of proteins secreted by the type III secretion apparatus of *Shigella flexneri*.** *Molecular microbiology* 2000, **38**:760-771.
276. Hale TL: **Genetic basis of virulence in *Shigella* species.** *Microbiol Rev* 1991, **55**:206-224.
277. Hartman AB, Venkatesan M, Oaks EV, Buysse JM: **Sequence and molecular characterization of a multicopy invasion plasmid antigen gene, *ipaH*, of *Shigella flexneri*.** *Journal of bacteriology* 1990, **172**:1905-1915.
278. Jin Q, Yuan Z, Xu J, Wang Y, Shen Y, Lu W, Wang J, Liu H, Yang J, Yang F, et al.: **Genome sequence of *Shigella flexneri* 2a: insights into pathogenicity through comparison with genomes of *Escherichia coli* K12 and O157.** *Nucleic acids research* 2002, **30**:4432-4441.
279. Yang F, Yang J, Zhang X, Chen L, Jiang Y, Yan Y, Tang X, Wang J, Xiong Z, Dong J, et al.: **Genome dynamics and diversity of *Shigella* species, the etiologic agents of bacillary dysentery.** *Nucleic acids research* 2005, **33**:6445-6458.
280. Buysse JM, Hartman AB, Strockbine N, Venkatesan M: **Genetic polymorphism of the *ipaH* multicopy antigen gene in *Shigella* spp. and enteroinvasive *Escherichia coli*.** *Microbial pathogenesis* 1995, **19**:335-349.
281. Escobar-Páramo P, Giudicelli C, Parsot C, Denamur E: **The evolutionary history of *Shigella* and enteroinvasive *Escherichia coli* revised.** *Journal of molecular evolution* 2003, **57**:140-148.
282. Belotserkovsky I, Sansonetti PJ: ***Shigella* and Enteroinvasive *Escherichia Coli*.** *Curr Top Microbiol Immunol* 2018, **416**:1-26.
283. The HC, Thanh DP, Holt KE, Thomson NR, Baker S: **The genomic signatures of *Shigella* evolution, adaptation and geographical spread.** *Nature reviews Microbiology* 2016, **14**:235-250.
284. Hazen TH, Leonard SR, Lampel KA, Lacher DW, Maurelli AT, Rasko DA: **Investigating the Relatedness of Enteroinvasive *Escherichia coli* to Other *E*.**

- coli* and *Shigella* Isolates by Using Comparative Genomics. *Infection and immunity* 2016, **84**:2362-2371.
285. Unkmeir A, Schmidt H: **Structural analysis of phage-borne stx genes and their flanking sequences in shiga toxin-producing *Escherichia coli* and *Shigella dysenteriae* type 1 strains.** *Infect Immun* 2000, **68**:4856-4864.
  286. Beutin L, Strauch E, Fischer I: **Isolation of *Shigella sonnei* lysogenic for a bacteriophage encoding gene for production of Shiga toxin.** *Lancet (London, England)* 1999, **353**:1498-1498.
  287. Gray MD, Lampel KA, Strockbine NA, Fernandez RE, Melton-Celsa AR, Maurelli AT: **Clinical isolates of Shiga toxin 1a-producing *Shigella flexneri* with an epidemiological link to recent travel to Hispaniola.** *Emerg Infect Dis* 2014, **20**:1669-1677.
  288. Gray MD, Leonard SR, Lacher DW, Lampel KA, Alam MT, Morris JG, Jr., Ali A, LaBreck PT, Maurelli AT: **Stx-Producing *Shigella* Species From Patients in Haiti: An Emerging Pathogen With the Potential for Global Spread.** *Open Forum Infect Dis* 2015, **2**:ofv134.
  289. Gupta SK, Strockbine N, Omondi M, Hise K, Fair MA, Mintz E: **Emergence of Shiga toxin 1 genes within *Shigella dysenteriae* type 4 isolates from travelers returning from the Island of Hispaniola.** *Am J Trop Med Hyg* 2007, **76**:1163-1165.
  290. Nyholm O, Lienemann T, Halkilahti J, Mero S, Rimhanen-Finne R, Lehtinen V, Salmenlinna S, Siitonen A: **Characterization of *Shigella sonnei* Isolate Carrying Shiga Toxin 2-Producing Gene.** *Emerging infectious diseases* 2015, **21**:891-892.
  291. Pupo GM, Lan R, Reeves PR: **Multiple independent origins of *Shigella* clones of *Escherichia coli* and convergent evolution of many of their characteristics.** *Proc Natl Acad Sci U S A* 2000, **97**:10567-10572.
  292. Yang J, Nie H, Chen L, Zhang X, Yang F, Xu X, Zhu Y, Yu J, Jin Q: **Revisiting the molecular evolutionary history of *Shigella* spp.** *Journal of molecular evolution* 2007, **64**:71-79.
  293. Brenner DJ, Steigerwalt AG, Wathen HG, Gross RJ, Rowe B: **Confirmation of aerogenic strains of *Shigella boydii* 13 and further study of *Shigella* serotypes by DNA relatedness.** *J Clin Microbiol* 1982, **16**:432-436.

294. Hyma KE, Lacher DW, Nelson AM, Bumbaugh AC, Janda JM, Strockbine NA, Young VB, Whittam TS: **Evolutionary genetics of a new pathogenic *Escherichia* species: *Escherichia albertii* and related *Shigella boydii* strains.** *Journal of bacteriology* 2005, **187**:619-628.
295. Lan R, Reeves PR: ***Escherichia coli* in disguise: molecular origins of *Shigella*.** *Microbes Infect* 2002, **4**:1125-1132.
296. Pettengill EA, Pettengill JB, Binet R: **Phylogenetic Analyses of *Shigella* and Enteroinvasive *Escherichia coli* for the Identification of Molecular Epidemiological Markers: Whole-Genome Comparative Analysis Does Not Support Distinct Genera Designation.** *Frontiers in microbiology* 2015, **6**:1573.
297. Niyogi SK: **Shigellosis.** *J Microbiol* 2005, **43**:133-143.
298. Lindsay B, Ochieng JB, Ikumapayi UN, Toure A, Ahmed D, Li S, Panchalingam S, Levine MM, Kotloff K, Rasko DA, et al.: **Quantitative PCR for detection of *Shigella* improves ascertainment of *Shigella* burden in children with moderate-to-severe diarrhea in low-income countries.** *Journal of clinical microbiology* 2013, **51**:1740-1746.
299. Song T, Toma C, Nakasone N, Iwanaga M: **Sensitive and rapid detection of *Shigella* and enteroinvasive *Escherichia coli* by a loop-mediated isothermal amplification method.** *FEMS Microbiol Lett* 2005, **243**:259-263.
300. Venkatesan MM, Buysse JM, Kopecko DJ: **Use of *Shigella flexneri ipaC* and *ipaH* gene sequences for the general identification of *Shigella* spp. and enteroinvasive *Escherichia coli*.** *Journal of clinical microbiology* 1989, **27**:2687-2691.
301. Vu DT, Sethabutr O, Von Seidlein L, Tran VT, Do GC, Bui TC, Le HT, Lee H, Hounh HS, Hale TL, et al.: **Detection of *Shigella* by a PCR assay targeting the *ipaH* gene suggests increased prevalence of shigellosis in Nha Trang, Vietnam.** *Journal of clinical microbiology* 2004, **42**:2031-2035.
302. Dhakal R, Wang Q, Lan R, Howard P, Sintchenko V: **Novel multiplex PCR assay for identification and subtyping of enteroinvasive *Escherichia coli* and differentiation from *Shigella* based on target genes selected by comparative genomics.** *Journal of medical microbiology* 2018, **67**:1257-1264.



303. Cunningham SA, Sloan LM, Nyre LM, Vetter EA, Mandrekar J, Patel R: **Three-hour molecular detection of *Campylobacter*, *Salmonella*, *Yersinia*, and *Shigella* species in feces with accuracy as high as that of culture.** *Journal of clinical microbiology* 2010, **48**:2929-2933.
304. Løbersli I, Wester AL, Kristiansen Å, Brandal LT: **Molecular Differentiation of *Shigella* Spp. from Enteroinvasive *E. Coli*.** *Eur J Microbiol Immunol (Bp)* 2016, **6**:197-205.
305. Pavlovic M, Luze A, Konrad R, Berger A, Sing A, Busch U, Huber I: **Development of a duplex real-time PCR for differentiation between *E. coli* and *Shigella* spp.** *Journal of applied microbiology* 2011, **110**:1245-1251.
306. Pettengill EA, Hoffmann M, Binet R, Roberts RJ, Payne J, Allard M, Michelacci V, Minelli F, Morabito S: **Complete Genome Sequence of Enteroinvasive *Escherichia coli* O96:H19 Associated with a Severe Foodborne Outbreak.** *Genome announcements* 2015, **3**.
307. Ito H, Kido N, Arakawa Y, Ohta M, Sugiyama T, Kato N: **Possible mechanisms underlying the slow lactose fermentation phenotype in *Shigella* spp.** *Appl Environ Microbiol* 1991, **57**:2912-2917.
308. Farfán MJ, Garay TA, Prado CA, Filliol I, Ulloa MT, Toro CS: **A new multiplex PCR for differential identification of *Shigella flexneri* and *Shigella sonnei* and detection of *Shigella* virulence determinants.** *Epidemiol Infect* 2010, **138**:525-533.
309. Kim HJ, Ryu JO, Song JY, Kim HY: **Multiplex Polymerase Chain Reaction for Identification of Shigellae and Four *Shigella* Species Using Novel Genetic Markers Screened by Comparative Genomics.** *Foodborne pathogens and disease* 2017, **14**:400-406.
310. Chattaway MA, Schaefer U, Tewolde R, Dallman TJ, Jenkins C: **Identification of *Escherichia coli* and *Shigella* Species from Whole-Genome Sequences.** *Journal of clinical microbiology* 2017, **55**:616-623.
311. Carlin NI, Lindberg AA: **Monoclonal antibodies specific for O-antigenic polysaccharides of *Shigella flexneri*: clones binding to II, II:3,4, and 7,8 epitopes.** *J Clin Microbiol* 1983, **18**:1183-1189.

312. Lefebvre J, Gosselin F, Ismail J, Lorange M, Lior H, Woodward D: **Evaluation of commercial antisera for *Shigella* serogrouping.** *J Clin Microbiol* 1995, **33**:1997-2001.
313. DebRoy C, Roberts E, Fratamico PM: **Detection of O antigens in *Escherichia coli*.** *Anim Health Res Rev* 2011, **12**:169-185.
314. Li Y, Cao B, Liu B, Liu D, Gao Q, Peng X, Wu J, Bastin DA, Feng L, Wang L: **Molecular detection of all 34 distinct O-antigen forms of *Shigella*.** *Journal of medical microbiology* 2009, **58**:69-81.
315. Coimbra RS, Grimont F, Grimont PA: **Identification of *Shigella* serotypes by restriction of amplified O-antigen gene cluster.** *Research in microbiology* 1999, **150**:543-553.
316. Coimbra RS, Lenormand P, Grimont F, Bouvet P, Matsushita S, Grimont PA: **Molecular and phenotypic characterization of potentially new *Shigella dysenteriae* serotype.** *J Clin Microbiol* 2001, **39**:618-621.
317. Grimont F, Lejay-Collin M, Talukder KA, Carle I, Issenhuth S, Le Roux K, Grimont PAD: **Identification of a group of shigella-like isolates as *Shigella boydii* 20.** *J Med Microbiol* 2007, **56**:749-754.
318. Melito PL, Woodward DL, Munro J, Walsh J, Foster R, Tilley P, Paccagnella A, Isaac-Renton J, Ismail J, Ng LK: **A novel *Shigella dysenteriae* serovar isolated in Canada.** *J Clin Microbiol* 2005, **43**:740-744.
319. de Paula CM, Mercedes PG, do Amaral PH, Tondo EC: **Antimicrobial resistance and PCR-ribotyping of *Shigella* responsible for foodborne outbreaks occurred in southern Brazil.** *Braz J Microbiol* 2010, **41**:966-977.
320. Coimbra RS, Artiguenave F, Jacques LS, Oliveira GC: **MST (molecular serotyping tool): a program for computer-assisted molecular identification of *Escherichia coli* and *Shigella* O antigens.** *J Clin Microbiol* 2010, **48**:1921-1923.
321. Sun Q, Lan R, Wang Y, Zhao A, Zhang S, Wang J, Wang Y, Xia S, Jin D, Cui Z, et al.: **Development of a multiplex PCR assay targeting O-antigen modification genes for molecular serotyping of *Shigella flexneri*.** *Journal of clinical microbiology* 2011, **49**:3766-3770.

322. van der Ploeg CA, Rogé AD, Bordagorria XL, de Urquiza MT, Castillo ABC, Bruno SB: **Design of Two Multiplex PCR Assays for Serotyping *Shigella flexneri*.** *Foodborne pathogens and disease* 2018, **15**:33-38.
323. Iguchi A, Iyoda S, Seto K, Morita-Ishihara T, Scheutz F, Ohnishi M: ***Escherichia coli* O-Genotyping PCR: a Comprehensive and Practical Platform for Molecular O Serogrouping.** *J Clin Microbiol* 2015, **53**:2427-2432.
324. Liu Y, Fratamico P: ***Escherichia coli* O antigen typing using DNA microarrays.** *Mol Cell Probes* 2006, **20**:239-244.
325. Coimbra RS, Grimont F, Lenormand P, Burguière P, Beutin L, Grimont PA: **Identification of *Escherichia coli* O-serogroups by restriction of the amplified O-antigen gene cluster (*rfb*-RFLP).** *Res Microbiol* 2000, **151**:639-654.
326. Wu Y, Lau HK, Lee T, Lau DK, Payne J: ***in Silico* Serotyping Based on Whole-Genome Sequencing Improves the Accuracy of *Shigella* Identification.** *Applied and environmental microbiology* 2019, **85**.
327. Casalino M, Latella MC, Prosseda G, Ceccarini P, Grimont F, Colonna B: **Molecular evolution of the lysine decarboxylase-defective phenotype in *Shigella sonnei*.** *Int J Med Microbiol* 2005, **294**:503-512.
328. Day WA, Jr., Fernández RE, Maurelli AT: **Pathoadaptive mutations that enhance virulence: genetic organization of the *cadA* regions of *Shigella* spp.** *Infect Immun* 2001, **69**:7471-7480.
329. Joensen KG, Tetzschner AM, Iguchi A, Aarestrup FM, Scheutz F: **Rapid and Easy *In Silico* Serotyping of *Escherichia coli* Isolates by Use of Whole-Genome Sequencing Data.** *Journal of clinical microbiology* 2015, **53**:2410-2426.
330. González-Escalona N, Kase JA: **Virulence gene profiles and phylogeny of Shiga toxin-positive *Escherichia coli* strains isolated from FDA regulated foods during 2010-2017.** *PLoS One* 2019, **14**:e0214620.
331. World Health Organization: *Shiga Toxin-producing Escherichia Coli (STEC) and Food: Attribution Characterization and Monitoring*, vol 19. Edited by Organization WH: World Health Organization; 2019.
332. Majowicz SE, Scallan E, Jones-Bitton A, Sargeant JM, Stapleton J, Angulo FJ, Yeung DH, Kirk MD: **Global incidence of human Shiga toxin-producing**

- Escherichia coli* infections and deaths: a systematic review and knowledge synthesis.** *Foodborne Pathog Dis* 2014, **11**:447-455.
333. Riley LW, Remis RS, Helgerson SD, McGee HB, Wells JG, Davis BR, Hebert RJ, Olcott ES, Johnson LM, Hargrett NT, et al.: **Hemorrhagic colitis associated with a rare *Escherichia coli* serotype.** *N Engl J Med* 1983, **308**:681-685.
334. Gyles CL: **Shiga toxin-producing *Escherichia coli*: an overview.** *J Anim Sci* 2007, **85**:E45-62.
335. Bettelheim KA, Goldwater PN: **Serotypes of non-O157 Shigatoxigenic *Escherichia coli* (STEC).** *J Advances in Microbiology* 2014, **2014**.
336. Karmali MA, Mascarenhas M, Shen S, Ziebell K, Johnson S, Reid-Smith R, Isaac-Renton J, Clark C, Rahn K, Kaper JB: **Association of genomic O island 122 of *Escherichia coli* EDL 933 with verocytotoxin-producing *Escherichia coli* seropathotypes that are linked to epidemic and/or serious disease.** *Journal of clinical microbiology* 2003, **41**:4930-4940.
337. Thorpe CM: **Shiga toxin-producing *Escherichia coli* infection.** *Clin Infect Dis* 2004, **38**:1298-1303.
338. Buvens G, De Gheldre Y, Dediste A, de Moreau AI, Mascart G, Simon A, Allemeersch D, Scheutz F, Lauwers S, Piérard D: **Incidence and virulence determinants of verocytotoxin-producing *Escherichia coli* infections in the Brussels-Capital Region, Belgium, in 2008-2010.** *J Clin Microbiol* 2012, **50**:1336-1345.
339. Rivas M, Miliwebsky E, Chinen I, Roldán CD, Balbi L, García B, Fiorilli G, Sosa-Estani S, Kincaid J, Rangel J, et al.: **Characterization and epidemiologic subtyping of Shiga toxin-producing *Escherichia coli* strains isolated from hemolytic uremic syndrome and diarrhea cases in Argentina.** *Foodborne Pathog Dis* 2006, **3**:88-96.
340. Vally H, Hall G, Dyda A, Raupach J, Knope K, Combs B, Desmarchelier P: **Epidemiology of Shiga toxin producing *Escherichia coli* in Australia, 2000-2010.** *BMC Public Health* 2012, **12**:63.
341. Mead PS, Griffin PM: ***Escherichia coli* O157:H7.** *Lancet* 1998, **352**:1207-1212.
342. Reiss G, Kunz P, Koin D, Keeffe EB: ***Escherichia coli* O157:H7 infection in nursing homes: review of literature and report of recent outbreak.** *J Am Geriatr Soc* 2006, **54**:680-684.

343. Tarr PI, Gordon CA, Chandler WL: **Shiga-toxin-producing *Escherichia coli* and haemolytic uraemic syndrome.** *Lancet* 2005, **365**:1073-1086.
344. Kobayashi N, Lee K, Yamazaki A, Saito S, Furukawa I, Kono T, Maeda E, Isobe J, Sugita-Konishi Y, Hara-Kudo Y: **Virulence gene profiles and population genetic analysis for exploration of pathogenic serogroups of Shiga toxin-producing *Escherichia coli*.** *J Clin Microbiol* 2013, **51**:4022-4028.
345. Armstrong GL, Hollingsworth J, Morris JG, Jr.: **Emerging foodborne pathogens: *Escherichia coli* O157:H7 as a model of entry of a new pathogen into the food supply of the developed world.** *Epidemiol Rev* 1996, **18**:29-51.
346. Gould LH, Mody RK, Ong KL, Clogher P, Cronquist AB, Garman KN, Lathrop S, Medus C, Spina NL, Webb TH, et al.: **Increased recognition of non-O157 Shiga toxin-producing *Escherichia coli* infections in the United States during 2000-2010: epidemiologic features and comparison with *E. coli* O157 infections.** *Foodborne Pathog Dis* 2013, **10**:453-460.
347. Control CfD, Prevention: **National Shiga toxin-producing *Escherichia coli* (STEC) surveillance annual report, 2016.** *J Centers for Disease Control Prevention, Atlanta* 2018.
348. Blankenship HM, Mosci RE, Dietrich S, Burgess E, Wholehan J, McWilliams K, Pietrzen K, Benko S, Gatesy T, Rudrik JT, et al.: **Population structure and genetic diversity of non-O157 Shiga toxin-producing *Escherichia coli* (STEC) clinical isolates from Michigan.** *Sci Rep* 2021, **11**:4461.
349. Brooks JT, Sowers EG, Wells JG, Greene KD, Griffin PM, Hoekstra RM, Strockbine NA: **Non-O157 Shiga toxin-producing *Escherichia coli* infections in the United States, 1983-2002.** *The Journal of infectious diseases* 2005, **192**:1422-1429.
350. Panel EB, Koutsoumanis K, Allende A, Alvarez-Ordóñez A, Bover-Cid S, Chemaly M, Davies R, De Cesare A, Herman L, Hilbert F: **Pathogenicity assessment of Shiga toxin-producing *Escherichia coli* (STEC) and the public health risk posed by contamination of food with STEC.** *J EFSA Journal* 2020, **18**:e05967.
351. Johnson WM, Lior H, Bezanson GS: **Cytotoxic *Escherichia coli* O157:H7 associated with haemorrhagic colitis in Canada.** *Lancet* 1983, **1**:76.

352. Heiman KE, Mody RK, Johnson SD, Griffin PM, Gould LH: ***Escherichia coli* O157 Outbreaks in the United States, 2003-2012.** *Emerg Infect Dis* 2015, **21**:1293-1301.
353. Vasant BR, Stafford RJ, Jennison AV, Bennett SM, Bell RJ, Doyle CJ, Young JR, Vlack SA, Titmus P, El Saadi D, et al.: **Mild Illness during Outbreak of Shiga Toxin-Producing *Escherichia coli* O157 Infections Associated with Agricultural Show, Australia.** *Emerg Infect Dis* 2017, **23**:1686-1689.
354. Kuehne A, Bouwknegt M, Havelaar A, Gilsdorf A, Hoyer P, Stark K, Werber D: **Estimating true incidence of O157 and non-O157 Shiga toxin-producing *Escherichia coli* illness in Germany based on notification data of haemolytic uraemic syndrome.** *Epidemiol Infect* 2016, **144**:3305-3315.
355. Mylius M, Dreesman J, Pulz M, Pallasch G, Beyrer K, Claußen K, Allerberger F, Fruth A, Lang C, Prager R, et al.: **Shiga toxin-producing *Escherichia coli* O103:H2 outbreak in Germany after school trip to Austria due to raw cow milk, 2017 - The important role of international collaboration for outbreak investigations.** *Int J Med Microbiol* 2018, **308**:539-544.
356. Rasko DA, Webster DR, Sahl JW, Bashir A, Boisen N, Scheutz F, Paxinos EE, Sebra R, Chin CS, Iliopoulos D, et al.: **Origins of the *E. coli* strain causing an outbreak of hemolytic-uremic syndrome in Germany.** *N Engl J Med* 2011, **365**:709-717.
357. Frank C, Werber D, Cramer JP, Askar M, Faber M, an der Heiden M, Bernard H, Fruth A, Prager R, Spode A, et al.: **Epidemic profile of Shiga-toxin-producing *Escherichia coli* O104:H4 outbreak in Germany.** *N Engl J Med* 2011, **365**:1771-1780.
358. King LA, Nogareda F, Weill FX, Mariani-Kurkdjian P, Loukiadis E, Gault G, Jourdan-DaSilva N, Bingen E, Macé M, Thevenot D, et al.: **Outbreak of Shiga toxin-producing *Escherichia coli* O104:H4 associated with organic fenugreek sprouts, France, June 2011.** *Clin Infect Dis* 2012, **54**:1588-1594.
359. Soysal N, Mariani-Kurkdjian P, Smail Y, Liguori S, Gouali M, Loukiadis E, Fach P, Bruyand M, Blanco J, Bidet P, et al.: **Enterohemorrhagic *Escherichia coli* Hybrid Pathotype O80:H2 as a New Therapeutic Challenge.** *Emerg Infect Dis* 2016, **22**:1604-1612.

360. Persad AK, LeJeune JT: **Animal Reservoirs of Shiga Toxin-Producing *Escherichia coli*.** *Microbiol Spectr* 2014, **2**:Ehec-0027-2014.
361. Herman KM, Hall AJ, Gould LH: **Outbreaks attributed to fresh leafy vegetables, United States, 1973-2012.** *Epidemiol Infect* 2015, **143**:3011-3021.
362. Busani L, Boccia D, Caprioli A, F MR, Morabito S, Minelli F, Lana S, Rizzoni G, Giofrè F, Mazzeo M, et al.: **Public health implications of a case of haemolytic-uraemic syndrome associated with a concomitant outbreak of mild gastroenteritis in a small rural community.** *Epidemiol Infect* 2006, **134**:407-413.
363. O'Brien AD, Tesh VL, Donohue-Rolfe A, Jackson MP, Olsnes S, Sandvig K, Lindberg AA, Keusch GT: **Shiga toxin: biochemistry, genetics, mode of action, and role in pathogenesis.** *Curr Top Microbiol Immunol* 1992, **180**:65-94.
364. Scheutz F, Teel LD, Beutin L, Piérard D, Buvens G, Karch H, Mellmann A, Caprioli A, Tozzoli R, Morabito S, et al.: **Multicenter evaluation of a sequence-based protocol for subtyping Shiga toxins and standardizing Stx nomenclature.** *J Clin Microbiol* 2012, **50**:2951-2963.
365. Yang X, Bai X, Zhang J, Sun H, Fu S, Fan R, He X, Scheutz F, Matussek A, Xiong Y: ***Escherichia coli* strains producing a novel Shiga toxin 2 subtype circulate in China.** *Int J Med Microbiol* 2020, **310**:151377.
366. Bai X, Fu S, Zhang J, Fan R, Xu Y, Sun H, He X, Xu J, Xiong Y: **Identification and pathogenomic analysis of an *Escherichia coli* strain producing a novel Shiga toxin 2 subtype.** *Sci Rep* 2018, **8**:6756.
367. Lacher DW, Gangiredla J, Patel I, Elkins CA, Feng PC: **Use of the *Escherichia coli* Identification Microarray for Characterizing the Health Risks of Shiga Toxin-Producing *Escherichia coli* Isolated from Foods.** *J Food Prot* 2016, **79**:1656-1662.
368. Melton-Celsa AR: **Shiga Toxin (Stx) Classification, Structure, and Function.** *Microbiol Spectr* 2014, **2**:Ehec-0024-2013.
369. Krüger A, Lucchesi PM: **Shiga toxins and stx phages: highly diverse entities.** *Microbiology (Reading)* 2015, **161**:451-462.
370. Bergan J, Dyve Lingelem AB, Simm R, Skotland T, Sandvig K: **Shiga toxins.** *Toxicon* 2012, **60**:1085-1107.

371. Johannes L, Römer W: **Shiga toxins--from cell biology to biomedical applications.** *Nat Rev Microbiol* 2010, **8**:105-116.
372. Tesh VL, O'Brien AD: **The pathogenic mechanisms of Shiga toxin and the Shiga-like toxins.** *Mol Microbiol* 1991, **5**:1817-1822.
373. O'Brien AD, Newland JW, Miller SF, Holmes RK, Smith HW, Formal SB: **Shiga-like toxin-converting phages from *Escherichia coli* strains that cause hemorrhagic colitis or infantile diarrhea.** *Science* 1984, **226**:694-696.
374. Scotland SM, Smith HR, Willshaw GA, Rowe B: **Vero cytotoxin production in strain of *Escherichia coli* is determined by genes carried on bacteriophage.** *Lancet* 1983, **2**:216.
375. Smith HW, Green P, Parsell Z: **Vero cell toxins in *Escherichia coli* and related bacteria: transfer by phage and conjugation and toxic action in laboratory animals, chickens and pigs.** *J Gen Microbiol* 1983, **129**:3121-3137.
376. Neely MN, Friedman DI: **Functional and genetic analysis of regulatory regions of coliphage H-19B: location of shiga-like toxin and lysis genes suggest a role for phage functions in toxin release.** *Mol Microbiol* 1998, **28**:1255-1267.
377. Friedrich AW, Bielaszewska M, Zhang WL, Pulz M, Kuczius T, Ammon A, Karch H: ***Escherichia coli* harboring Shiga toxin 2 gene variants: frequency and association with clinical symptoms.** *J Infect Dis* 2002, **185**:74-84.
378. Luna-Gierke RE, Griffin PM, Gould LH, Herman K, Bopp CA, Strockbine N, Mody RK: **Outbreaks of non-O157 Shiga toxin-producing *Escherichia coli* infection: USA.** *Epidemiol Infect* 2014, **142**:2270-2280.
379. Shringi S, García A, Lahmers KK, Potter KA, Muthupalani S, Swennes AG, Hovde CJ, Call DR, Fox JG, Besser TE: **Differential virulence of clinical and bovine-biased enterohemorrhagic *Escherichia coli* O157:H7 genotypes in piglet and Dutch belted rabbit models.** *Infect Immun* 2012, **80**:369-380.
380. Boerlin P, McEwen SA, Boerlin-Petzold F, Wilson JB, Johnson RP, Gyles CL: **Associations between virulence factors of Shiga toxin-producing *Escherichia coli* and disease in humans.** *J Clin Microbiol* 1999, **37**:497-503.
381. Bielaszewska M, Friedrich AW, Aldick T, Schürk-Bulgrin R, Karch H: **Shiga toxin activatable by intestinal mucus in *Escherichia coli* isolated from humans: predictor for a severe clinical outcome.** *Clin Infect Dis* 2006, **43**:1160-1167.



382. De Rauw K, Jacobs S, Piérard D: **Twenty-seven years of screening for Shiga toxin-producing *Escherichia coli* in a university hospital. Brussels, Belgium, 1987-2014.** *PLoS One* 2018, **13**:e0199968.
383. Etoh Y, Murakami K, Ichihara S, Sera N, Hamasaki M, Takenaka S, Horikawa K, Kawano K, Takeishi T, Kuwana Y, et al.: **Isolation of Shiga toxin 2f-producing *Escherichia coli* (O115:HNM) from an adult symptomatic patient in Fukuoka Prefecture, Japan.** *Jpn J Infect Dis* 2009, **62**:315-317.
384. Sonntag AK, Zenner E, Karch H, Bielaszewska M: **Pigeons as a possible reservoir of Shiga toxin 2f-producing *Escherichia coli* pathogenic to humans.** *Berl Munch Tierarztl Wochenschr* 2005, **118**:464-470.
385. Seto K, Taguchi M, Kobayashi K, Kozaki S: **Biochemical and molecular characterization of minor serogroups of Shiga toxin-producing *Escherichia coli* isolated from humans in Osaka prefecture.** *J Vet Med Sci* 2007, **69**:1215-1222.
386. Fasel D, Mellmann A, Cernela N, Hächler H, Fruth A, Khanna N, Egli A, Beckmann C, Hirsch HH, Goldenberger D, et al.: **Hemolytic uremic syndrome in a 65-Year-old male linked to a very unusual type of *stx2e*- and *eae*-harboring O51:H49 shiga toxin-producing *Escherichia coli*.** *J Clin Microbiol* 2014, **52**:1301-1303.
387. Gray MD, Lacher DW, Leonard SR, Abbott J, Zhao S, Lampel KA, Prothery E, Gouali M, Weill FX, Maurelli AT: **Prevalence of Shiga toxin-producing *Shigella* species isolated from French travellers returning from the Caribbean: an emerging pathogen with international implications.** *Clin Microbiol Infect* 2015, **21**:765.e769-765.e714.
388. Blanco M, Blanco JE, Mora A, Dahbi G, Alonso MP, González EA, Bernárdez MI, Blanco J: **Serotypes, virulence genes, and intimin types of Shiga toxin (verotoxin)-producing *Escherichia coli* isolates from cattle in Spain and identification of a new intimin variant gene (*eae-xi*).** *Journal of clinical microbiology* 2004, **42**:645-651.
389. Lai Y, Rosenshine I, Leong JM, Frankel G: **Intimate host attachment: enteropathogenic and enterohaemorrhagic *Escherichia coli*.** *Cell Microbiol* 2013, **15**:1796-1808.

390. Adu-Bobie J, Frankel G, Bain C, Goncalves AG, Trabulsi LR, Douce G, Knutton S, Dougan G: **Detection of intimins alpha, beta, gamma, and delta, four intimin derivatives expressed by attaching and effacing microbial pathogens.** *J Clin Microbiol* 1998, **36**:662-668.
391. McGraw EA, Li J, Selander RK, Whittam TS: **Molecular evolution and mosaic structure of alpha, beta, and gamma intimins of pathogenic *Escherichia coli*.** *Mol Biol Evol* 1999, **16**:12-22.
392. Jerse AE, Yu J, Tall BD, Kaper JB: **A genetic locus of enteropathogenic *Escherichia coli* necessary for the production of attaching and effacing lesions on tissue culture cells.** *Proc Natl Acad Sci U S A* 1990, **87**:7839-7843.
393. Beutin L, Zimmermann S, Gleier K: **Rapid detection and isolation of shiga-like toxin (verocytotoxin)-producing *Escherichia coli* by direct testing of individual enterohemolytic colonies from washed sheep blood agar plates in the VTEC-RPLA assay.** *J Clin Microbiol* 1996, **34**:2812-2814.
394. Schmidt H, Beutin L, Karch H: **Molecular analysis of the plasmid-encoded hemolysin of *Escherichia coli* O157:H7 strain EDL 933.** *Infect Immun* 1995, **63**:1055-1061.
395. Bauer ME, Welch RA: **Characterization of an RTX toxin from enterohemorrhagic *Escherichia coli* O157:H7.** *Infect Immun* 1996, **64**:167-175.
396. Lim JY, Yoon J, Hovde CJ: **A brief overview of *Escherichia coli* O157:H7 and its plasmid O157.** *J Microbiol Biotechnol* 2010, **20**:5-14.
397. Schmidt H, Karch H: **Enterohemolytic phenotypes and genotypes of shiga toxin-producing *Escherichia coli* O111 strains from patients with diarrhea and hemolytic-uremic syndrome.** *J Clin Microbiol* 1996, **34**:2364-2367.
398. Asadulghani M, Ogura Y, Ooka T, Itoh T, Sawaguchi A, Iguchi A, Nakayama K, Hayashi T: **The defective prophage pool of *Escherichia coli* O157: prophage-prophage interactions potentiate horizontal transfer of virulence determinants.** *PLoS Pathog* 2009, **5**:e1000408.
399. Herold S, Karch H, Schmidt H: **Shiga toxin-encoding bacteriophages--genomes in motion.** *Int J Med Microbiol* 2004, **294**:115-121.

400. Smith DL, Rooks DJ, Fogg PC, Darby AC, Thomson NR, McCarthy AJ, Allison HE: **Comparative genomics of Shiga toxin encoding bacteriophages.** *BMC Genomics* 2012, **13**:311.
401. Muniesa M, de Simon M, Prats G, Ferrer D, Pañella H, Jofre J: **Shiga toxin 2-converting bacteriophages associated with clonal variability in *Escherichia coli* O157:H7 strains of human origin isolated from a single outbreak.** *Infect Immun* 2003, **71**:4554-4562.
402. Teel LD, Melton-Celsa AR, Schmitt CK, O'Brien AD: **One of two copies of the gene for the activatable shiga toxin type 2d in *Escherichia coli* O91:H21 strain B2F1 is associated with an inducible bacteriophage.** *Infect Immun* 2002, **70**:4282-4291.
403. Schmidt H: **Shiga-toxin-converting bacteriophages.** *Res Microbiol* 2001, **152**:687-695.
404. Ferdous M, Zhou K, Mellmann A, Morabito S, Croughs PD, de Boer RF, Kooistra-Smid AM, Rossen JW, Friedrich AW: **Is Shiga Toxin-Negative *Escherichia coli* O157:H7 Enteropathogenic or Enterohemorrhagic *Escherichia coli*? Comprehensive Molecular Analysis Using Whole-Genome Sequencing.** *J Clin Microbiol* 2015, **53**:3530-3538.
405. Mora A, López C, Dhabí G, López-Beceiro AM, Fidalgo LE, Díaz EA, Martínez-Carrasco C, Mamani R, Herrera A, Blanco JE, et al.: **Seropathotypes, Phylogroups, Stx subtypes, and intimin types of wildlife-carried, shiga toxin-producing *escherichia coli* strains with the same characteristics as human-pathogenic isolates.** *Applied and environmental microbiology* 2012, **78**:2578-2585.
406. Bielaszewska M, Köck R, Friedrich AW, von Eiff C, Zimmerhackl LB, Karch H, Mellmann A: **Shiga toxin-mediated hemolytic uremic syndrome: time to change the diagnostic paradigm?** *PLoS One* 2007, **2**:e1024.
407. McDaniel TK, Jarvis KG, Donnenberg MS, Kaper JB: **A genetic locus of enterocyte effacement conserved among diverse enterobacterial pathogens.** *Proc Natl Acad Sci U S A* 1995, **92**:1664-1668.
408. Hacker J, Kaper JB: **Pathogenicity islands and the evolution of microbes.** *Annu Rev Microbiol* 2000, **54**:641-679.

409. Deng W, Li Y, Vallance BA, Finlay BB: **Locus of enterocyte effacement from *Citrobacter rodentium*: sequence analysis and evidence for horizontal transfer among attaching and effacing pathogens.** *Infect Immun* 2001, **69**:6323-6335.
410. Schmidt H, Karch H, Beutin L: **The large-sized plasmids of enterohemorrhagic *Escherichia coli* O157 strains encode hemolysins which are presumably members of the *E. coli* alpha-hemolysin family.** *FEMS Microbiol Lett* 1994, **117**:189-196.
411. Schmidt H, Kernbach C, Karch H: **Analysis of the EHEC *hly* operon and its location in the physical map of the large plasmid of enterohaemorrhagic *Escherichia coli* O157:h7.** *Microbiology (Reading)* 1996, **142 ( Pt 4)**:907-914.
412. Brunder W, Schmidt H, Karch H: **KatP, a novel catalase-peroxidase encoded by the large plasmid of enterohaemorrhagic *Escherichia coli* O157:H7.** *Microbiology (Reading)* 1996, **142 ( Pt 11)**:3305-3315.
413. Schmidt H, Henkel B, Karch H: **A gene cluster closely related to type II secretion pathway operons of gram-negative bacteria is located on the large plasmid of enterohemorrhagic *Escherichia coli* O157 strains.** *FEMS Microbiol Lett* 1997, **148**:265-272.
414. Brunder W, Schmidt H, Karch H: **EspP, a novel extracellular serine protease of enterohaemorrhagic *Escherichia coli* O157:H7 cleaves human coagulation factor V.** *Mol Microbiol* 1997, **24**:767-778.
415. Tatsuno I, Horie M, Abe H, Miki T, Makino K, Shinagawa H, Taguchi H, Kamiya S, Hayashi T, Sasakawa C: ***tox*B gene on pO157 of enterohemorrhagic *Escherichia coli* O157:H7 is required for full epithelial cell adherence phenotype.** *Infect Immun* 2001, **69**:6660-6669.
416. Yoon JW, Lim JY, Park YH, Hovde CJ: **Involvement of the *Escherichia coli* O157:H7(pO157) ecf operon and lipid A myristoyl transferase activity in bacterial survival in the bovine gastrointestinal tract and bacterial persistence in farm water troughs.** *Infect Immun* 2005, **73**:2367-2378.
417. Melton-Celsa A, Mohawk K, Teel L, O'Brien A: **Pathogenesis of Shiga-toxin producing *escherichia coli*.** *Curr Top Microbiol Immunol* 2012, **357**:67-103.
418. Yoon JW, Hovde CJ: **All blood, no stool: enterohemorrhagic *Escherichia coli* O157:H7 infection.** *J Vet Sci* 2008, **9**:219-231.

419. Fratamico PM, Yan X, Caprioli A, Esposito G, Needleman DS, Pepe T, Tozzoli R, Cortesi ML, Morabito S: **The complete DNA sequence and analysis of the virulence plasmid and of five additional plasmids carried by Shiga toxin-producing *Escherichia coli* O26:H11 strain H30.** *Int J Med Microbiol* 2011, **301**:192-203.
420. Eichhorn I, Heidemanns K, Semmler T, Kinnemann B, Mellmann A, Harmsen D, Anjum MF, Schmidt H, Fruth A, Valentin-Weigand P, et al.: **Highly Virulent Non-O157 Enterohemorrhagic *Escherichia coli* (EHEC) Serotypes Reflect Similar Phylogenetic Lineages, Providing New Insights into the Evolution of EHEC.** *Appl Environ Microbiol* 2015, **81**:7041-7047.
421. Wick LM, Qi W, Lacher DW, Whittam TS: **Evolution of genomic content in the stepwise emergence of *Escherichia coli* O157:H7.** *J Bacteriol* 2005, **187**:1783-1791.
422. Bai X, Zhang J, Ambikan A, Jernberg C, Ehricht R, Scheutz F, Xiong Y, Matussek A: **Molecular Characterization and Comparative Genomics of Clinical Hybrid Shiga Toxin-Producing and Enterotoxigenic *Escherichia coli* (STEC/ETEC) Strains in Sweden.** *Sci Rep* 2019, **9**:5619.
423. March SB, Ratnam S: **Sorbitol-MacConkey medium for detection of *Escherichia coli* O157:H7 associated with hemorrhagic colitis.** *J Clin Microbiol* 1986, **23**:869-872.
424. Karch H, Wiss R, Gloning H, Emmrich P, Aleksić S, Bockemühl J: **[Hemolytic-uremic syndrome in infants due to verotoxin-producing *Escherichia coli*].** *Dtsch Med Wochenschr* 1990, **115**:489-495.
425. Rosser T, Dransfield T, Allison L, Hanson M, Holden N, Evans J, Naylor S, La Ragione R, Low JC, Gally DL: **Pathogenic potential of emergent sorbitol-fermenting *Escherichia coli* O157:NM.** *Infect Immun* 2008, **76**:5598-5607.
426. Bettelheim KA, Whipp M, Djordjevic SP, Ramachandran V: **First isolation outside Europe of sorbitol-fermenting verocytotoxigenic *Escherichia coli* (VTEC) belonging to O group O157.** *J Med Microbiol* 2002, **51**:713-714.
427. Friedrich AW, Zhang W, Bielaszewska M, Mellmann A, Köck R, Fruth A, Tschäpe H, Karch H: **Prevalence, virulence profiles, and clinical significance of Shiga toxin-negative variants of enterohemorrhagic *Escherichia coli* O157 infection in humans.** *Clin Infect Dis* 2007, **45**:39-45.

428. Pennington H: *Escherichia coli* O157. *Lancet* 2010, **376**:1428-1435.
429. Leopold SR, Magrini V, Holt NJ, Shaikh N, Mardis ER, Cagno J, Ogura Y, Iguchi A, Hayashi T, Mellmann A, et al.: **A precise reconstruction of the emergence and constrained radiations of *Escherichia coli* O157 portrayed by backbone concatenomic analysis.** *Proc Natl Acad Sci U S A* 2009, **106**:8713-8718.
430. Shaikh N, Tarr PI: ***Escherichia coli* O157:H7 Shiga toxin-encoding bacteriophages: integrations, excisions, truncations, and evolutionary implications.** *J Bacteriol* 2003, **185**:3596-3605.
431. Frank C, Faber MS, Askar M, Bernard H, Fruth A, Gilsdorf A, Hohle M, Karch H, Krause G, Prager R, et al.: **Large and ongoing outbreak of haemolytic uraemic syndrome, Germany, May 2011.** *Euro Surveill* 2011, **16**.
432. Bielaszewska M, Mellmann A, Zhang W, Köck R, Fruth A, Bauwens A, Peters G, Karch H: **Characterisation of the *Escherichia coli* strain associated with an outbreak of haemolytic uraemic syndrome in Germany, 2011: a microbiological study.** *Lancet Infect Dis* 2011, **11**:671-676.
433. Navarro-Garcia F: ***Escherichia coli* O104:H4 Pathogenesis: an Enteraggregative *E. coli*/Shiga Toxin-Producing *E. coli* Explosive Cocktail of High Virulence.** *Microbiol Spectr* 2014, **2**.
434. Mariani-Kurkdjian P, Lemaître C, Bidet P, Perez D, Boggini L, Kwon T, Bonacorsi S: **Haemolytic-uraemic syndrome with bacteraemia caused by a new hybrid *Escherichia coli* pathotype.** *New Microbes New Infect* 2014, **2**:127-131.
435. De Rauw K, Thiry D, Caljon B, Saulmont M, Mainil J, Piérard D: **Characteristics of Shiga toxin producing- and enteropathogenic *Escherichia coli* of the emerging serotype O80:H2 isolated from humans and diarrhoeic calves in Belgium.** *Clin Microbiol Infect* 2019, **25**:111.e115-111.e118.
436. Nüesch-Inderbilen M, Cernela N, Wüthrich D, Egli A, Stephan R: **Genetic characterization of Shiga toxin producing *Escherichia coli* belonging to the emerging hybrid pathotype O80:H2 isolated from humans 2010-2017 in Switzerland.** *Int J Med Microbiol* 2018, **308**:534-538.
437. Nyholm O, Heinikainen S, Pelkonen S, Hallanvuo S, Haukka K, Siitonen A: **Hybrids of Shigatoxigenic and Enterotoxigenic *Escherichia coli***

- (STEC/ETEC) Among Human and Animal Isolates in Finland. *Zoonoses Public Health* 2015, **62**:518-524.
438. Oh KH, Shin E, Jung SM, Im J, Cho SH, Hong S, Yoo CK, Chung GT: **First Isolation of a Hybrid Shigatoxigenic and Enterotoxigenic *Escherichia coli* Strain Harboring the *stx2* and *elt* Genes in Korea.** *Jpn J Infect Dis* 2017, **70**:347-348.
439. Zadik PM, Chapman PA, Siddons CA: **Use of tellurite for the selection of verocytotoxigenic *Escherichia coli* O157.** *J Med Microbiol* 1993, **39**:155-158.
440. Gunzer F, Böhm H, Rüssmann H, Bitzan M, Aleksic S, Karch H: **Molecular detection of sorbitol-fermenting *Escherichia coli* O157 in patients with hemolytic-uremic syndrome.** *J Clin Microbiol* 1992, **30**:1807-1810.
441. Bettelheim KA: **Reliability of CHROMagar O157 for the detection of enterohaemorrhagic *Escherichia coli* (EHEC) O157 but not EHEC belonging to other serogroups.** *J Appl Microbiol* 1998, **85**:425-428.
442. Zelyas N, Poon A, Patterson-Fortin L, Johnson RP, Lee W, Chui L: **Assessment of commercial chromogenic solid media for the detection of non-O157 Shiga toxin-producing *Escherichia coli* (STEC).** *Diagn Microbiol Infect Dis* 2016, **85**:302-308.
443. Wylie JL, Van Caesele P, Gilmour MW, Sitter D, Guttek C, Giercke S: **Evaluation of a new chromogenic agar medium for detection of Shiga toxin-producing *Escherichia coli* (STEC) and relative prevalences of O157 and non-O157 STEC in Manitoba, Canada.** *J Clin Microbiol* 2013, **51**:466-471.
444. Hirvonen JJ, Siitonen A, Kaukoranta SS: **Usability and performance of CHROMagar STEC medium in detection of Shiga toxin-producing *Escherichia coli* strains.** *J Clin Microbiol* 2012, **50**:3586-3590.
445. Milley DG, Sekla LH: **An enzyme-linked immunosorbent assay-based isolation procedure for verotoxigenic *Escherichia coli*.** *Appl Environ Microbiol* 1993, **59**:4223-4229.
446. Perera LP, Marques LR, O'Brien AD: **Isolation and characterization of monoclonal antibodies to Shiga-like toxin II of enterohemorrhagic *Escherichia coli* and use of the monoclonal antibodies in a colony enzyme-linked immunosorbent assay.** *J Clin Microbiol* 1988, **26**:2127-2131.

447. Teel LD, Daly JA, Jerris RC, Maul D, Svanas G, O'Brien AD, Park CH: **Rapid detection of Shiga toxin-producing *Escherichia coli* by optical immunoassay.** *J Clin Microbiol* 2007, **45**:3377-3380.
448. Orskov I, Orskov F, Jann B, Jann K: **Serology, chemistry, and genetics of O and K antigens of *Escherichia coli*.** *Bacteriol Rev* 1977, **41**:667-710.
449. Stigi KA, Macdonald JK, Tellez-Marfin AA, Lofy KH: **Laboratory practices and incidence of non-O157 shiga toxin-producing *Escherichia coli* infections.** *Emerg Infect Dis* 2012, **18**:477-479.
450. Pollard DR, Johnson WM, Lior H, Tyler SD, Rozee KR: **Rapid and specific detection of verotoxin genes in *Escherichia coli* by the polymerase chain reaction.** *J Clin Microbiol* 1990, **28**:540-545.
451. Scotland SM, Rowe B, Smith HR, Willshaw GA, Gross RJ: **Vero cytotoxin-producing strains of *Escherichia coli* from children with haemolytic uraemic syndrome and their detection by specific DNA probes.** *J Med Microbiol* 1988, **25**:237-243.
452. Bélanger SD, Boissinot M, Ménard C, Picard FJ, Bergeron MG: **Rapid detection of Shiga toxin-producing bacteria in feces by multiplex PCR with molecular beacons on the smart cycler.** *J Clin Microbiol* 2002, **40**:1436-1440.
453. Grys TE, Sloan LM, Rosenblatt JE, Patel R: **Rapid and sensitive detection of Shiga toxin-producing *Escherichia coli* from nonenriched stool specimens by real-time PCR in comparison to enzyme immunoassay and culture.** *J Clin Microbiol* 2009, **47**:2008-2012.
454. Gerritzen A, Wittke JW, Wolff D: **Rapid and sensitive detection of Shiga toxin-producing *Escherichia coli* directly from stool samples by real-time PCR in comparison to culture, enzyme immunoassay and Vero cell cytotoxicity assay.** *Clin Lab* 2011, **57**:993-998.
455. Zhang W, Bielaszewska M, Bauwens A, Fruth A, Mellmann A, Karch H: **Real-time multiplex PCR for detecting Shiga toxin 2-producing *Escherichia coli* O104:H4 in human stools.** *J Clin Microbiol* 2012, **50**:1752-1754.
456. Chui L, Couturier MR, Chiu T, Wang G, Olson AB, McDonald RR, Antonishyn NA, Horsman G, Gilmour MW: **Comparison of Shiga toxin-producing *Escherichia coli* detection methods using clinical stool samples.** *J Mol Diagn* 2010, **12**:469-475.



457. Parsons BD, Zelyas N, Berenger BM, Chui L: **Detection, Characterization, and Typing of Shiga Toxin-Producing *Escherichia coli***. *Front Microbiol* 2016, **7**:478.
458. DebRoy C, Roberts E, Davis M, Bumbaugh A: **Multiplex polymerase chain reaction assay for detection of nonserotypable Shiga toxin-producing *Escherichia coli* strains of serogroup O147**. *Foodborne Pathog Dis* 2010, **7**:1407-1414.
459. DebRoy C, Roberts E, Valadez AM, Dudley EG, Cutter CN: **Detection of Shiga toxin-producing *Escherichia coli* O26, O45, O103, O111, O113, O121, O145, and O157 serogroups by multiplex polymerase chain reaction of the *wzx* gene of the O-antigen gene cluster**. *Foodborne Pathog Dis* 2011, **8**:651-652.
460. DebRoy C, Roberts E, Kundra J, Davis MA, Briggs CE, Fratamico PM: **Detection of *Escherichia coli* serogroups O26 and O113 by PCR amplification of the *wzx* and *wzy* genes**. *Appl Environ Microbiol* 2004, **70**:1830-1832.
461. Feng L, Senchenkova SN, Tao J, Shashkov AS, Liu B, Shevelev SD, Reeves PR, Xu J, Knirel YA, Wang L: **Structural and genetic characterization of enterohemorrhagic *Escherichia coli* O145 O antigen and development of an O145 serogroup-specific PCR assay**. *J Bacteriol* 2005, **187**:758-764.
462. Lin A, Sultan O, Lau HK, Wong E, Hartman G, Lauzon CR: **O serogroup specific real time PCR assays for the detection and identification of nine clinically relevant non-O157 STECs**. *Food Microbiol* 2011, **28**:478-483.
463. Fratamico PM, Bagi LK: **Detection of Shiga toxin-producing *Escherichia coli* in ground beef using the GeneDisc real-time PCR system**. *Front Cell Infect Microbiol* 2012, **2**:152.
464. Beutin L, Jahn S, Fach P: **Evaluation of the 'GeneDisc' real-time PCR system for detection of enterohaemorrhagic *Escherichia coli* (EHEC) O26, O103, O111, O145 and O157 strains according to their virulence markers and their O- and H-antigen-associated genes**. *J Appl Microbiol* 2009, **106**:1122-1132.
465. Fratamico PM, DebRoy C, Strobaugh TP, Jr., Chen CY: **DNA sequence of the *Escherichia coli* O103 O antigen gene cluster and detection of enterohemorrhagic *E. coli* O103 by PCR amplification of the *wzx* and *wzy* genes**. *Can J Microbiol* 2005, **51**:515-522.

466. Ludwig JB, Shi X, Shridhar PB, Roberts EL, DebRoy C, Phebus RK, Bai J, Nagaraja TG: **Multiplex PCR Assays for the Detection of One Hundred and Thirty Seven Serogroups of Shiga Toxin-Producing *Escherichia coli* Associated With Cattle.** *Front Cell Infect Microbiol* 2020, **10**:378.
467. Sánchez S, Llorente MT, Echeita MA, Herrera-León S: **Development of three multiplex PCR assays targeting the 21 most clinically relevant serogroups associated with Shiga toxin-producing *E. coli* infection in humans.** *PLoS One* 2015, **10**:e0117660.
468. Shridhar PB, Noll LW, Shi X, An B, Cernicchiaro N, Renter DG, Nagaraja TG, Bai J: **Multiplex Quantitative PCR Assays for the Detection and Quantification of the Six Major Non-O157 *Escherichia coli* Serogroups in Cattle Feces.** *J Food Prot* 2016, **79**:66-74.
469. Pintara AP, Guglielmino CJD, Rathnayake IU, Huygens F, Jennison AV: **Molecular Prediction of the O157:H-Negative Phenotype Prevalent in Australian Shiga Toxin-Producing *Escherichia coli* Cases Improves Concordance of *In Silico* Serotyping with Phenotypic Motility.** *J Clin Microbiol* 2018, **56**.
470. Ribot EM, Fair MA, Gautom R, Cameron DN, Hunter SB, Swaminathan B, Barrett TJ: **Standardization of pulsed-field gel electrophoresis protocols for the subtyping of *Escherichia coli* O157:H7, *Salmonella*, and *Shigella* for PulseNet.** *Foodborne Pathog Dis* 2006, **3**:59-67.
471. Swaminathan B, Gerner-Smidt P, Ng LK, Lukinmaa S, Kam KM, Rolando S, Gutiérrez EP, Binsztein N: **Building PulseNet International: an interconnected system of laboratory networks to facilitate timely public health recognition and response to foodborne disease outbreaks and emerging foodborne diseases.** *Foodborne Pathog Dis* 2006, **3**:36-50.
472. Lindstedt BA, Vardund T, Kapperud G: **Multiple-Locus Variable-Number Tandem-Repeats Analysis of *Escherichia coli* O157 using PCR multiplexing and multi-colored capillary electrophoresis.** *J Microbiol Methods* 2004, **58**:213-222.
473. Hyytiä-Trees E, Smole SC, Fields PA, Swaminathan B, Ribot EM: **Second generation subtyping: a proposed PulseNet protocol for multiple-locus**

- variable-number tandem repeat analysis of Shiga toxin-producing *Escherichia coli* O157 (STEC O157).** *Foodborne Pathog Dis* 2006, **3**:118-131.
474. Jolley KA, Maiden MC: **BIGSdb: Scalable analysis of bacterial genome variation at the population level.** *BMC Bioinformatics* 2010, **11**:595.
475. Amézquita-López BA, Quiñones B, Cooley MB, León-Félix J, Castro-del Campo N, Mandrell RE, Jiménez M, Chaidez C: **Genotypic analyses of shiga toxin-producing *Escherichia coli* O157 and non-O157 recovered from feces of domestic animals on rural farms in Mexico.** *PLoS One* 2012, **7**:e51565.
476. Dallman TJ, Byrne L, Launders N, Glen K, Grant KA, Jenkins C: **The utility and public health implications of PCR and whole genome sequencing for the detection and investigation of an outbreak of Shiga toxin-producing *Escherichia coli* serogroup O26:H11.** *Epidemiol Infect* 2015, **143**:1672-1680.
477. Jenkins C, Dallman TJ, Launders N, Willis C, Byrne L, Jorgensen F, Eppinger M, Adak GK, Aird H, Elviss N, et al.: **Public Health Investigation of Two Outbreaks of Shiga Toxin-Producing *Escherichia coli* O157 Associated with Consumption of Watercress.** *Appl Environ Microbiol* 2015, **81**:3946-3952.
478. Holmes A, Allison L, Ward M, Dallman TJ, Clark R, Fawkes A, Murphy L, Hanson M: **Utility of Whole-Genome Sequencing of *Escherichia coli* O157 for Outbreak Detection and Epidemiological Surveillance.** *J Clin Microbiol* 2015, **53**:3565-3573.
479. Chattaway MA, Dallman TJ, Gentle A, Wright MJ, Long SE, Ashton PM, Perry NT, Jenkins C: **Whole Genome Sequencing for Public Health Surveillance of Shiga Toxin-Producing *Escherichia coli* Other than Serogroup O157.** *Front Microbiol* 2016, **7**:258.
480. Dallman TJ, Byrne L, Ashton PM, Cowley LA, Perry NT, Adak G, Petrovska L, Ellis RJ, Elson R, Underwood A, et al.: **Whole-genome sequencing for national surveillance of Shiga toxin-producing *Escherichia coli* O157.** *Clin Infect Dis* 2015, **61**:305-312.
481. Gilchrist CA, Turner SD, Riley MF, Petri WA, Jr., Hewlett EL: **Whole-genome sequencing in outbreak analysis.** *Clin Microbiol Rev* 2015, **28**:541-563.
482. Sabat AJ, Budimir A, Nashev D, Sá-Leão R, van Dijl J, Laurent F, Grundmann H, Friedrich AW: **Overview of molecular typing methods for outbreak detection and epidemiological surveillance.** *Euro Surveill* 2013, **18**:20380.

483. Saltykova A, Buytaers FE, Denayer S, Verhaegen B, Piérard D, Roosens NHC, Marchal K, De Keersmaecker SCJ: **Strain-Level Metagenomic Data Analysis of Enriched In Vitro and In Silico Spiked Food Samples: Paving the Way towards a Culture-Free Foodborne Outbreak Investigation Using STEC as a Case Study.** *Int J Mol Sci* 2020, **21**.
484. Buytaers FE, Saltykova A, Denayer S, Verhaegen B, Vanneste K, Roosens NHC, Piérard D, Marchal K, De Keersmaecker SCJ: **A Practical Method to Implement Strain-Level Metagenomics-Based Foodborne Outbreak Investigation and Source Tracking in Routine.** *Microorganisms* 2020, **8**.
485. Koutsoumanis K, Allende A, Alvarez-Ordóñez A, Bolton D, Bover-Cid S, Chemaly M, Davies R, De Cesare A, Hilbert F, Lindqvist R, et al.: **Whole genome sequencing and metagenomics for outbreak investigation, source attribution and risk assessment of food-borne microorganisms.** *Efsa j* 2019, **17**:e05898.
486. Loman NJ, Constantinidou C, Christner M, Rohde H, Chan JZ, Quick J, Weir JC, Quince C, Smith GP, Betley JR, et al.: **A culture-independent sequence-based metagenomics approach to the investigation of an outbreak of Shiga-toxigenic *Escherichia coli* O104:H4.** *Jama* 2013, **309**:1502-1510.
487. Huang AD, Luo C, Pena-Gonzalez A, Weigand MR, Tarr CL, Konstantinidis KT: **Metagenomics of Two Severe Foodborne Outbreaks Provides Diagnostic Signatures and Signs of Coinfection Not Attainable by Traditional Methods.** *Appl Environ Microbiol* 2017, **83**.
488. Leonard SR, Mammel MK, Lacher DW, Elkins CA: **Application of metagenomic sequencing to food safety: detection of Shiga Toxin-producing *Escherichia coli* on fresh bagged spinach.** *Appl Environ Microbiol* 2015, **81**:8183-8191.
489. Leonard SR, Mammel MK, Lacher DW, Elkins CA: **Strain-Level Discrimination of Shiga Toxin-Producing *Escherichia coli* in Spinach Using Metagenomic Sequencing.** *PLoS One* 2016, **11**:e0167870.
490. McInerney JO, McNally A, O'Connell MJ: **Why prokaryotes have pangenomes.** *Nat Microbiol* 2017, **2**:17040.
491. Gordienko EN, Kazanov MD, Gelfand MS: **Evolution of pan-genomes of *Escherichia coli*, *Shigella* spp., and *Salmonella enterica*.** *J Bacteriol* 2013, **195**:2786-2792.

492. Kim HJ, Park SH, Lee TH, Nahm BH, Chung YH, Seo KH, Kim HY: **Identification of *Salmonella enterica* serovar Typhimurium using specific PCR primers obtained by comparative genomics in *Salmonella* serovars.** *J Food Prot* 2006, **69**:1653-1661.
493. Kim HJ, Park SH, Kim HY: **Comparison of *Salmonella enterica* serovar Typhimurium LT2 and non-LT2 *Salmonella* genomic sequences, and genotyping of salmonellae by using PCR.** *Appl Environ Microbiol* 2006, **72**:6142-6151.
494. Shi C, Singh P, Ranieri ML, Wiedmann M, Moreno Switt AI: **Molecular methods for serovar determination of *Salmonella*.** *Crit Rev Microbiol* 2015, **41**:309-325.
495. Zou QH, Li RQ, Wang YJ, Liu SL: **Identification of genes to differentiate closely related *Salmonella* lineages.** *PLoS One* 2013, **8**:e55988.
496. Ford L, Moffatt CR, Fearnley E, Miller M, Gregory J, Sloan-Gardner TS, Polkinghorne BG, Bell R, Franklin N, Williamson DA: **The epidemiology of *Salmonella enterica* outbreaks in Australia, 2001–2016.** *Frontiers in Sustainable Food Systems* 2018, **2**:86.
497. Group. OW: **Monitoring the incidence and causes of diseases potentially transmitted by food in Australia: annual report of the OzFoodNet Network, 2009.** *Communicable diseases intelligence quarterly report* 2010, **34**:396-426.
498. Group. OW: **Monitoring the incidence and causes of diseases potentially transmitted by food in Australia: annual report of the OzFoodNet network, 2010.** *Communicable diseases intelligence quarterly report* 2012, **36**:E213-241.
499. Bettelheim KA: **Role of non-O157 VTEC.** *Symp Ser Soc Appl Microbiol* 2000:38s-50s.
500. Li B, Liu H, Wang W: **Multiplex real-time PCR assay for detection of *Escherichia coli* O157:H7 and screening for non-O157 Shiga toxin-producing *E. coli*.** *BMC Microbiol* 2017, **17**:215.
501. Qin X, Klein EJ, Galanakis E, Thomas AA, Stapp JR, Rich S, Buccat AM, Tarr PI: **Real-Time PCR Assay for Detection and Differentiation of Shiga Toxin-Producing *Escherichia coli* from Clinical Samples.** *J Clin Microbiol* 2015, **53**:2148-2153.

502. Paton AW, Woodrow MC, Doyle RM, Lanser JA, Paton JC: **Molecular characterization of a Shiga toxigenic *Escherichia coli* O113:H21 strain lacking eae responsible for a cluster of cases of hemolytic-uremic syndrome.** *J Clin Microbiol* 1999, **37**:3357-3361.
503. Zhang W, Mellmann A, Sonntag AK, Wieler L, Bielaszewska M, Tschäpe H, Karch H, Friedrich AW: **Structural and functional differences between disease-associated genes of enterohaemorrhagic *Escherichia coli* O111.** *Int J Med Microbiol* 2007, **297**:17-26.
504. McCarthy TA, Barrett NL, Hadler JL, Salsbury B, Howard RT, Dingman DW, Brinkman CD, Bibb WF, Cartter ML: **Hemolytic-Uremic Syndrome and *Escherichia coli* O121 at a Lake in Connecticut, 1999.** *Pediatrics* 2001, **108**:E59.
505. Morton V, Cheng JM, Sharma D, Kearney A: **Notes from the Field: An Outbreak of Shiga Toxin-Producing *Escherichia coli* O121 Infections Associated with Flour - Canada, 2016-2017.** *MMWR Morb Mortal Wkly Rep* 2017, **66**:705-706.
506. Johnson KE, Thorpe CM, Sears CL: **The emerging clinical importance of non-O157 Shiga toxin-producing *Escherichia coli*.** *Clin Infect Dis* 2006, **43**:1587-1595.
507. Käppeli U, Hächler H, Giezendanner N, Beutin L, Stephan R: **Human infections with non-O157 Shiga toxin-producing *Escherichia coli*, Switzerland, 2000-2009.** *Emerg Infect Dis* 2011, **17**:180-185.
508. Liptáková A, Siegfried L, Kmetová M, Birosová E, Kotulová D, Bencátová A, Kosecká M, Bánovcin P: **Hemolytic uremic syndrome caused by verotoxin-producing *Escherichia coli* O26. Case report.** *Folia Microbiol (Praha)* 2005, **50**:95-98.
509. Paciorek J: **Virulence properties of *Escherichia coli* faecal strains isolated in Poland from healthy children and strains belonging to serogroups O18, O26, O44, O86, O126 and O127 isolated from children with diarrhoea.** *J Med Microbiol* 2002, **51**:548-571.
510. Verstraete K, K DER, S VANW, Piérard D, L DEZ, Herman L, Robyn J, Heyndrickx M: **Genetic characteristics of Shiga toxin-producing *E. coli***

- O157, O26, O103, O111 and O145 isolates from humans, food, and cattle in Belgium.** *Epidemiol Infect* 2013, **141**:2503-2515.
511. Zweifel C, Cernela N, Stephan R: **Detection of the emerging Shiga toxin-producing *Escherichia coli* O26:H11/H- sequence type 29 (ST29) clone in human patients and healthy cattle in Switzerland.** *Appl Environ Microbiol* 2013, **79**:5411-5413.
512. European Food Safety Authority ECfDPC: **The European Union Summary Report on Trends and Sources of Zoonoses, Zoonotic Agents and Food-borne Outbreaks in 2009.** *EFSA Journal: European Food Standards Agency* 2011, **9**:2090.
513. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, et al.: **SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing.** *J Comput Biol* 2012, **19**:455-477.
514. Gurevich A, Saveliev V, Vyahhi N, Tesler G: **QUAST: quality assessment tool for genome assemblies.** *Bioinformatics* 2013, **29**:1072-1075.
515. Chen Z, Zhang K, Yin H, Li Q, Wang L, Liu Z: **Detection of *Salmonella* and several common *Salmonella* serotypes in food by loop-mediated isothermal amplification method.** *Food Science Human Wellness* 2015, **4**:75-79.
516. Shanmugasundaram M, Radhika M, Murali H, Batra H: **Detection of *Salmonella enterica* serovar Typhimurium by selective amplification of *fliC*, *fliB*, *iroB*, *invA*, *rfbJ*, STM2755, STM4497 genes by polymerase chain reaction in a monoplex and multiplex format.** *World Journal of Microbiology Biotechnology* 2009, **25**:1385-1394.
517. Ju W, Cao G, Rump L, Strain E, Luo Y, Timme R, Allard M, Zhao S, Brown E, Meng J: **Phylogenetic analysis of non-O157 Shiga toxin-producing *Escherichia coli* strains by whole-genome sequencing.** *J Clin Microbiol* 2012, **50**:4123-4127.
518. Hendriksen RS, Vieira AR, Karlsmose S, Lo Fo Wong DM, Jensen AB, Wegener HC, Aarestrup FM: **Global monitoring of *Salmonella* serovar distribution from the World Health Organization Global Foodborne Infections Network Country Data Bank: results of quality assured laboratories from 2001 to 2007.** *Foodborne Pathog Dis* 2011, **8**:887-900.

519. de Boer RF, Ott A, Kesztyüs B, Kooistra-Smid AM: **Improved detection of five major gastrointestinal pathogens by use of a molecular screening approach.** *Journal of clinical microbiology* 2010, **48**:4140-4146.
520. van den Beld MJC, Friedrich AW, van Zanten E, Reubsaet FAG, Kooistra-Smid M, Rossen JWA: **Multicenter evaluation of molecular and culture-dependent diagnostics for *Shigella* species and Enteroinvasive *Escherichia coli* in the Netherlands.** *Journal of microbiological methods* 2016, **131**:10-15.
521. Van Lint P, De Witte E, Ursi J, Van Herendael B, Van Schaeren J: **A screening algorithm for diagnosing bacterial gastroenteritis by real-time PCR in combination with guided culture.** *Diagnostic microbiology* 2016, **85**:255-259.
522. GROUP. FWSE: **Hazard Identification and Characterization: Criteria for Categorizing Shiga Toxin-Producing *Escherichia coli* on a Risk Basis(†).** *Journal of food protection* 2019, **82**:7-21.



# Appendix

## Appendix I: Supplementary Material of Chapter 2

**FIGURE S1** | The SNP based phylogenetic tree constructed by ParSNP showing the evolutionary relationships within and between serovars using 1344 representative isolates including 1258 isolates from 107 serovars examined in the study and 86 isolates from serovars with less than 5 rSTs which were otherwise excluded from the study.

**TABLE S1** | The final data set of 2258 high quality and consistent serovar prediction genomes representing 107 serovars.

**TABLE S2** | A total of 414 candidate serovar-specific genes including 295 serovar-specific genes and 119 lineage-specific genes.

**TABLE S3** | An additional 1089 validation isolates with serovar prediction results by SISTR, SeqSero and serovar-specific gene markers.

**TABLE S4** | A minimum of 131 genes for identification of 106 serovars.

**TABLE S5** | A set of 65 genes for identification of 46 common serovars.

**DATA S1** | Sequences of 131 serovar-specific gene markers.

**The Supplementary Material for this article can be found online at:**

<https://www.frontiersin.org/articles/10.3389/fmicb.2019.00835/full#supplementary-material>;

[https://drive.google.com/drive/folders/1VkW2goYTdT\\_KYjlnEf4vCsnXCuW5iKt1?usp=sharing](https://drive.google.com/drive/folders/1VkW2goYTdT_KYjlnEf4vCsnXCuW5iKt1?usp=sharing)

## Appendix II: Supplementary Material of Chapter 3

Supplemental material for this article can be found at:

<https://doi.org/10.1016/j.jmoldx.2020.02.006>.

<https://drive.google.com/drive/folders/1DpbvFwt32VMbM38vcmnWGPocRo4hiZYQ?usp=sharing>

### Appendix II: Table S1: Bacterial strains used in this study

1, Bacterial strains used for sensitivity testing: 79		
Lab ACC	Serovar	Source
L2	Typhimurium	SARA2
L3	Typhimurium	SARA3
L4	Typhimurium	SARA4
L6	Typhimurium	SARA6
L7	Typhimurium	SARA7
L8	Typhimurium	SARA8
L9	Typhimurium	SARA9
L10	Typhimurium	SARA10
L12	Typhimurium	SARA12
L14	Typhimurium	SARA14
L138	Typhimurium	SARB66
L90	Enteritidis	SARB18
L2376	Enteritidis	NSW Enteric Reference Laboratory, NSW Health Pathology
L2377	Enteritidis	NSW Enteric Reference Laboratory, NSW Health Pathology
L2378	Enteritidis	NSW Enteric Reference Laboratory, NSW Health Pathology
L2379	Enteritidis	NSW Enteric Reference Laboratory, NSW Health Pathology
L2380	Enteritidis	NSW Enteric Reference Laboratory, NSW Health Pathology
L2350	Enteritidis	NSW Enteric Reference Laboratory, NSW Health Pathology
L2351	Enteritidis	NSW Enteric Reference Laboratory, NSW Health Pathology
L2352	Enteritidis	NSW Enteric Reference Laboratory, NSW Health Pathology
L2353	Enteritidis	NSW Enteric Reference Laboratory, NSW Health Pathology
L2354	Enteritidis	NSW Enteric Reference Laboratory, NSW Health Pathology
L2355	Enteritidis	NSW Enteric Reference Laboratory, NSW Health Pathology
L2356	Enteritidis	NSW Enteric Reference Laboratory, NSW Health Pathology
L2357	Enteritidis	NSW Enteric Reference Laboratory, NSW Health Pathology
L2358	Enteritidis	NSW Enteric Reference Laboratory, NSW Health Pathology
L2359	Enteritidis	NSW Enteric Reference Laboratory, NSW Health Pathology
L2360	Enteritidis	NSW Enteric Reference Laboratory, NSW Health Pathology
L2361	Enteritidis	NSW Enteric Reference Laboratory, NSW Health Pathology
L2362	Enteritidis	NSW Enteric Reference Laboratory, NSW Health Pathology
L2363	Enteritidis	NSW Enteric Reference Laboratory, NSW Health Pathology
L2364	Enteritidis	NSW Enteric Reference Laboratory, NSW Health Pathology
L2371	Virchow	NSW Enteric Reference Laboratory, NSW Health Pathology
L2372	Virchow	NSW Enteric Reference Laboratory, NSW Health Pathology

L2373	Virchow	NSW Enteric Reference Laboratory, NSW Health Pathology
L2375	Virchow	NSW Enteric Reference Laboratory, NSW Health Pathology
L2344	Virchow	NSW Enteric Reference Laboratory, NSW Health Pathology
L2345	Virchow	NSW Enteric Reference Laboratory, NSW Health Pathology
L2346	Virchow	NSW Enteric Reference Laboratory, NSW Health Pathology
L2347	Virchow	NSW Enteric Reference Laboratory, NSW Health Pathology
L2348	Virchow	NSW Enteric Reference Laboratory, NSW Health Pathology
L2349	Virchow	NSW Enteric Reference Laboratory, NSW Health Pathology
L22	Saintpaul	SARA22
L23	Saintpaul	SARA23
L24	Saintpaul	SARA24
L27	Saintpaul	SARA27
L28	Saintpaul	SARA28
L29	Saintpaul	SARA29
L127	Saintpaul	SARB55
L128	Saintpaul	SARB56
L2386	Saintapul	NSW Enteric Reference Laboratory, NSW Health Pathology
L2387	Saintapul	NSW Enteric Reference Laboratory, NSW Health Pathology
L2388	Saintapul	NSW Enteric Reference Laboratory, NSW Health Pathology
L2389	Saintapul	NSW Enteric Reference Laboratory, NSW Health Pathology
L2390	Saintapul	NSW Enteric Reference Laboratory, NSW Health Pathology
L2396	Saintapul	NSW Enteric Reference Laboratory, NSW Health Pathology
L2365	Saintapul	NSW Enteric Reference Laboratory, NSW Health Pathology
L2366	Saintapul	NSW Enteric Reference Laboratory, NSW Health Pathology
L2367	Saintapul	NSW Enteric Reference Laboratory, NSW Health Pathology
L2368	Saintapul	NSW Enteric Reference Laboratory, NSW Health Pathology
L2369	Saintapul	NSW Enteric Reference Laboratory, NSW Health Pathology
L2370	Saintapul	NSW Enteric Reference Laboratory, NSW Health Pathology
L98	Infantis	SARB26
L99	Infantis	SARB27
L2374	Infantis	NSW Enteric Reference Laboratory, NSW Health Pathology
L2381	Infantis	NSW Enteric Reference Laboratory, NSW Health Pathology
L2382	Infantis	NSW Enteric Reference Laboratory, NSW Health Pathology
L2384	Infantis	NSW Enteric Reference Laboratory, NSW Health Pathology
L2385	Infantis	NSW Enteric Reference Laboratory, NSW Health Pathology
L2337	Infantis	NSW Enteric Reference Laboratory, NSW Health Pathology
L2338	Infantis	NSW Enteric Reference Laboratory, NSW Health Pathology
L2339	Infantis	NSW Enteric Reference Laboratory, NSW Health Pathology
L2340	Infantis	NSW Enteric Reference Laboratory, NSW Health Pathology
L2341	Infantis	NSW Enteric Reference Laboratory, NSW Health Pathology
L2342	Infantis	NSW Enteric Reference Laboratory, NSW Health Pathology
L2343	Infantis	NSW Enteric Reference Laboratory, NSW Health Pathology
2, Bacterial strains used for specificity testing: 38		
L73	Agona	SARB1
L74	Anatum	SARB2
L75	Brandenburg	SARB3

L76	Choleraesuis	SARB4
L81	Derby	SARB9
L85	Dublin	SARB13
L93	Gallinarum	SARB21
L95	Heidelberg	SARB23
L97	Indiana	SARB25
L101	Miami	SARB29
L102	Montevideo	SARB30
L107	Muenchen	SARB35
L109	Newport	SARB37
L111	Panama	SARB39
L115	Paratyphi B	SARB43
L121	Paratyphi C	SARB49
L123	Pullorum	SARB51
L125	Reading	SARB53
L129	Schwarzengrund	SARB57
L131	Senftenberg	SARB59
L132	Stanley	SARB60
L133	Stanleyville	SARB61
L134	Thompson	SARB62
L135	Typhi	SARB63
M884	Moellerella	
M892	Pseudomonas	
M898	Yersinia	
M902	Citrobacter	
M903	Enterobacter	
M904	Klebsiella	
K12	Escherichia	
L1607	Vibrio	
L2376	Enteritidis	NSW Enteric Reference Laboratory, NSW Health Pathology
L2378	Enteritidis	NSW Enteric Reference Laboratory, NSW Health Pathology
L2371	Virchow	NSW Enteric Reference Laboratory, NSW Health Pathology
L127	Saintpaul	SARB55
L128	Saintpaul	SARB56
L98	Infantis	SARB26
L138	Typhimurium	SARB66

**Appendix II: Table S2: *in silico* sensitivity and specificity of the seven MCDA products**

Database from previous study: 2258 genomes *										
Serovar	Selected gene	MCDA products	No of genomes	Non-target genomes	TP	TN	FP	FN	Sensitivity	Specificity
Typhimurium	STM4494	223	214	2044	214	1904	140	0	100	93.2
Enteritidis-clade B	SEN1384	215	139	2119	134	2110	9	5	96.4	99.6
Enteritidis-clade A/C	R561_RS18155	259	26	2232	26	2206	26	0	100	98.8
Virchow	SEV_RS01820	230	39	2219	39	2219	0	0	100	100
Saintpaul-I	SESPA_RS08460	248	24	2234	24	2155	79	0	100	96.5
Saintpaul-II	SeSPB_A1749	211	5	2253	5	2253	0	0	100	100
Infantis	L287_RS37190	209	33	2225	33	2223	2	0	100	99.9

MCDA: Multiple Cross Displacement Amplification.

\*: Zhang, X., Payne, M., and Lan, R. (2019). *In silico* Identification of Serovar-Specific Genes for *Salmonella* Serotyping. 10(835). doi: 10.3389/fmicb.2019.00835.

Sensitivity TP/(TP+FN)  
 Specificity TN/(TN+FP)  
 TP True positive  
 TN True negative  
 FP False positive  
 FN False negative

## **Appendix II: Data S1: The sequences of seven serovar/lineage-specific gene markers**

### **>Typhimurium specific gene STM4494**

ATGAAAATAGCGGCGATTTATACGCAATCCGTCGGACCGCTGCCCCGACGGTGAAAT  
TCGTTTTGAAAACGACTGGTCAGGCGAGATAGAATCCAATGTCCTGATTACCGGGC  
CAAACGGCTGCGGTAAATCGACGCTGTTACGAGCCATTTCTTATTATGGCGCGCCT  
TTGGGCATTGGCTGGGCACGGGTACGCGCCTGAATATCAAAGATGAATCTTATACC  
TGGTTCTATCGCTGGGATGCCAGCTGCGCGATGATTCTCGATTCTTTTCGCCAAAA  
TCGGAGGACCAAATAGGGCTATTTCTTGGCTCAGAGGCATTTTTAGTACAACATAAA  
AGAGAAATACCCACAGGTCTACTGGCTCGGGGAAACGGTAAGCAGAACAAAGGGA  
ACAACGCCAGAGACGACCGTTTTTACGTCATCGGAACACTTTTTTCTTCCCTATAAG  
AACTGGTGGAGTCACTGGAGTAGCCAGTATCAACGGCTTGTAAGGGACCAAG  
TGTTGATATGCCTAATCTGGTTTACCTGGATGCCGAAGCGCGCCGCTGGGTTCGCCC  
GCAAAAAGATATCGGCAGCCTGTCCCCGGATGATTCAACTCAAGCCTGGCTGGCAA  
CCTATGAAGTGAACGATAACTGGAAAGGTCAGCTTGAATCGTCGCTCTTCAATATG  
AAGGCAACCATGCCCGGTGAATATCCTGAAATGATCGCCACGCTAAACCAGTTCTT  
CAGCGGTAAACGTATCGAAGCTGAAATTCATCCAGGACAGCGGCAACGTGTCCTAC  
TGGATTCAAGGAATGATCATTTCGTTAGATGCGCTTAGCTCCGGCGAACATCAGGTG  
CTGATTATGCTGTTTACGGTACAACGCTGGCTACAGCCCGGCGGTGTTGTAATCATT  
GATGAGCCGGATCTGCATCTGCATCCGTCCCTGATATCGCCGTTGCTGGCATCCATT  
GAGAACATCGTTGCCAGGAAAAATGGTCAGCTTGTGATGACTTCACACGCAACGGA  
TATCTGGCAACGTTATGACAACATGGGATTGCGGATTGATTAAACCGATGGCAAGG  
ATGCGGAAAATGGCCAGCGTTAA

### **>Enteritidis-clade B specific gene SEN1384**

ATGAACTCCGGCCTGATAACACTTACGGAGCTGAGGAGGATGACGGGGTTAACCAT  
TTATTCGACCCGCCACTACCTGGACAAGGCAGAACGTTGTGGGGATGTGTACCAGG  
CGGGCAGAAGAGGGGGGATTTTCCCGTCAGAAGAGGCTTATCGTGCCTGGAAGAA  
ACAGGCGAAAGTGGACGCTGACCTGATTTGGAAGCTGCCTGACGGTGAGGTACGTC  
GTTACGACAGGCACCACAACGTAATTTGTCGTGAGTGTCGTAAGCGAGTACATG  
CAGCGGGTACTGGCGTTTTATCGGGGAAACTTTCAGGAGGTGCTGTTGTGA

### **>Enteritidis-clade A/C specific gene R561\_RS18155**

ATGATTGAAAAATTGGTTGATATTACCCCCCAAATATATCTTTAAAAGGTAGTCA  
GATAATTGATTTTCATTATAGGGCAGGGAGTCTTGAGATTATAGTTACTCTTGATGG

AGTGAATTCGGATTTTCGATTTTTTTTTTGGATTGGACTCATTCATTTTCGTGTCACTGA  
TGAAGGTGATCTGTTGAAAATGTTGGGTGAGCAAAAAGGAAAAATGCGAGTAGGT  
ATTTATAAGGTTGAGGACTCATCTTATCTGGAATGGTTTAATGACCAGAGTTTTAAT  
ATACATGAAAAAGAGAAAATTATTCATTATTTGATTGTGACAGTAAATGATATCAT  
TGATGTTTTGTCCTCAGAGTCTCCAGTGATATCTAACTGTTCTAAATAA

**>Virchow specific gene SESV\_RS06060**

ATGTTAAAAACACACATGAATGCAACCGAGAATCATTTGGTTTCTATCTCACAGAT  
TCCTGCTAATGCTGGACATACATTACATAGAGGTACACCGAGAGAAGCGTTTATTA  
AAGAGTTTCTTCCGGGCACTTAAGCTCTAATGTGGCAATTGGTTCAGGGGAAATT  
ATAGATTCTAACTCTCAACCAAGAGTACAAAGAAATCAGTATGATATTGTCATCTA  
TAAAAACAATTATCCAAAATTAGATTTTGGCGGTGGAGTTAATGGTTTTTTAATTGA  
GTCAGTAATTGCTACAATAGAAGTAAAATCATTATTAGACCAATCTGCAATTGACC  
AGTCGGTTAAGGCAGCTCACAAACGCGAAAGTTTTAAACCCAAGCATAAATAAAAGT  
TTCAGCACTGGATGGGTGCCACCTAAAATTATAAATTATGTTGTGGCATATGATGG  
GCCAGCACAAATGAATACTGTGTATAACTGGATCTTAAATAGTCATCAAACGAATA  
GGATCCCCTTGCCATCATGGAATCAGCAAACGAAATATCAAACACCAGGTACAGCA  
CTTGATGGTGTGGTTTTATTAAATAAAGGATTTATAAAGCTTGATAATACACCATTA  
TCGTTAAACTCTTCTCAGCAATCAGGAACCTCATATAGTTGTTGACAGTAATGATGGT  
AATCTATTAATGATGTTTTTTGGCTTTACAGGAAGCGTGTGACAATATCCAAGGCGCT  
TGGTTGAATGCAGGACCATATGTAAGAAATGTCGGATTTAACAATGTAAGAATAAT  
ATAA

**>Saintpaul-I specific gene SESPA\_RS08460**

ATGCTCCCGAATCGAATGGTACTTAGCCGTCAGACTGAAGACCAGCTTAAAAAGCT  
AAAAGGATATACGGGGATTACACCCAATGTCGCGGCTCGGCTGGCATTTCCTCGCT  
CAGTGGAGAGTGAGTTTCGCTATTTCGCTGAGCGGGATAGTAAGAAGCTGGATGGC  
TCTCTGGTGCTGGATAAAATAACGTGGCTGGGGGAAACGCTGCAAGCTACGGAGCT  
GGTACTAAAGATGCTATATCCGCAGCTAGAGCAGAAAATACTAATTAAAGCATGGG  
CAGCACATGTTGATGATGGAATTGCTGCATTAAGGAATTATCGAAGCTTAAAAGAT  
TTTTCAAAGAATATATAG

**>Saintpaul-II specific gene SeSPB\_A1749**

ATGAAAAGAATTGCCATCGACATGGATGAAGTTATCGCCGATTTTAACTGCAAATT  
TATTGCTTCTTTTAACGCCGTTTTTCTGAAAACATCACGGTCGCCGATCTCGCCGG

AAAAACGGTTGAGCAGTTCCGTCCACAACTGCTGGCGGAGATGCGGGCAATGATTT  
GCGAGGACGGTTTCTTCCGCGATATGCCGGTGATCCCGGACAGTCAGAGGGTGTT  
GAAGCATTGCATAACCGATACGAAATATTTATTACCAGTGCCGCGATGGACTGGCC  
GGGCTCTTTTAATGCTAAGTATCACTGGCTCCAGGAAAATTTCCCCTTCATCAGCCC  
TAAAAACGTGGTGTTTTGTGGCAACAAAAGCATTGTTTATGCTGACTACCTGATCG  
ATGACACTCCGCGACATTTCTCACCTTCCAGGGTGAAGGGATCCTGTTTTCTGCCC  
CCCATAACCTTGATACGGAAGGTTACCGCAGGGTTAATAGCTGGCTGGATGTGGAA  
ACGCTCTTCCTTTCATAA

**>Infantis specific gene L287\_11788**

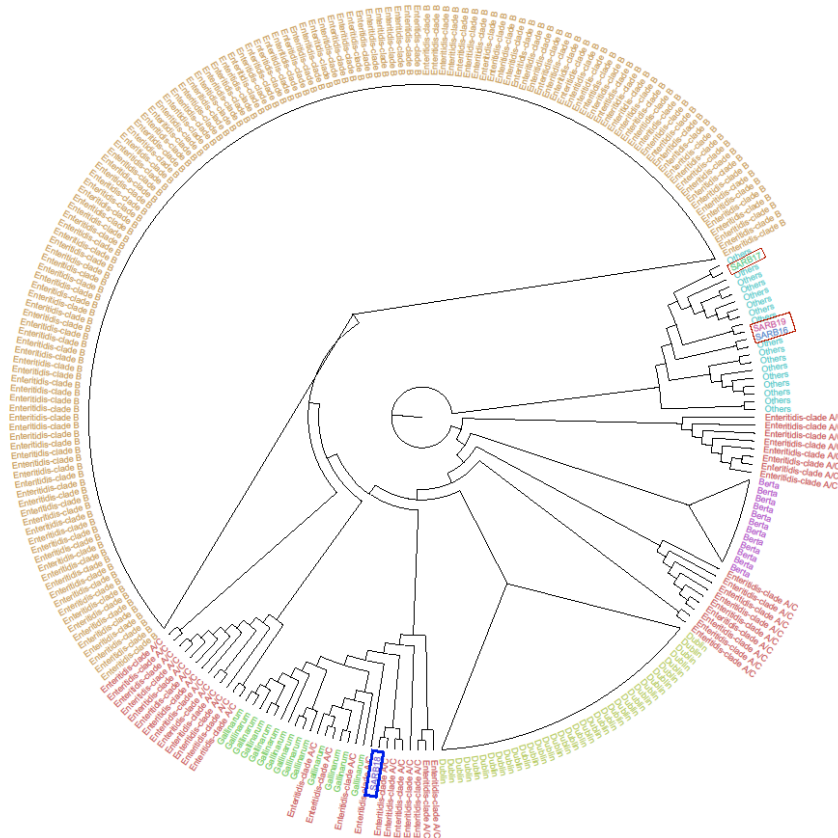
ATGTCAAATCAACCCAGTAACATCAGTAAAAACAGCGCTCTGCTGGTTATGGATTT  
TCAGACGATCATCCTTAACAATTTCTTCCCGCAAGAAAGCGCTGGAAACGTTATCC  
GCAATACCGCATCACTGATAGCCGCTGCGCGCACGGCAGGCGTACCGGTCATTTAT  
GTCAGTGTTCGGATTTTCGCGAGGGGTATCCAGAGGTCAGCAAAAACAACACTATCTT  
CTCTTCGATTAAAGAGAATGGAATTTTTATGGCTGACAACGAGAGCACGGCTATTC  
ATCCTGATGTCGCTCCGGCAGAAAATGAGGTCGTCATCGTTAAACGCAGAGTCGGA  
GCTTTCTCGTTTACCGAACTTGAGATGATCCTTCGTGCTCAAGGCATCGAAAACCTG  
ATCCTTACCGGTGTGACTACCAGTGGTGTTGTGCTTTCTACAGTCGGGCAAGCGTTT  
GACCTGGATTACCGCCTCATCGTTGTGAGTGATTGCTGCGCAGATCCAGATCATGA  
CACCAATGTGTTTTTACTTGAAAAAATTCTACCCCAACATGCTTTTGTTACCAGTTC  
ATCTGAAATATCAGAAGCCTGGGCATAA



**Appendix II: Figure S1: Phylogenetic relationship of 4 SARB Enteritidis strains.**

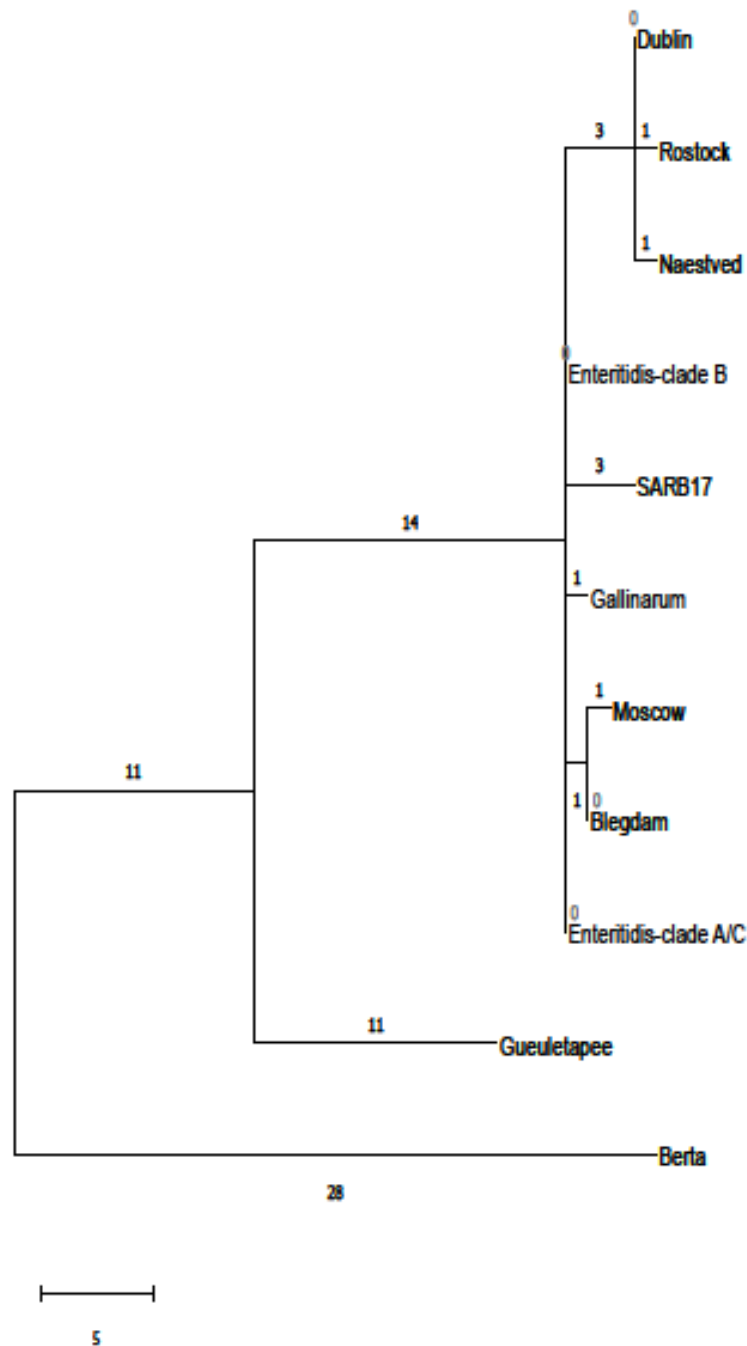
The SNP based phylogenetic tree showed the genetic relationship of 4 SARB Enteritidis strains with Enteritidis strains. SARB18 was grouped with Enteritidis-Clade A/C isolates, while SARB16, SARB17 and SARB19 were grouped with other serovars.

**SARB:** *Salmonella* Reference Collection B.



**Appendix II: Figure S2: Phylogenetic relationship of SARB17, Enteritidis and other serovars.**

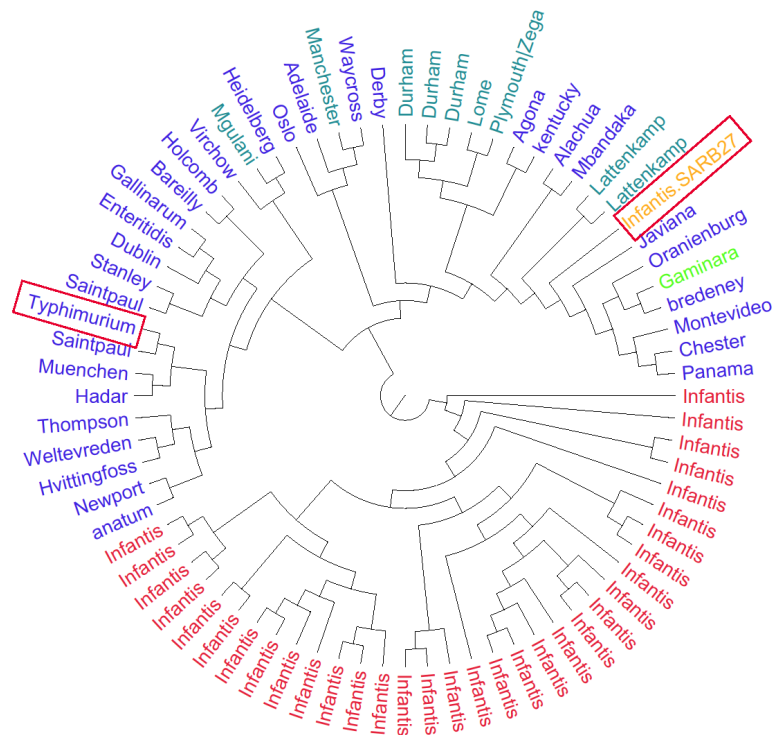
The SNP based phylogenetic tree constructed in MEGA X by Maximum Parsimony method for flagellin gene *fliC* of SARB17, Enteritidis-clade A/C, Enteritidis-Clade B, and other closely related to Enteritidis serovars. Tree length are indicated at the nodes. Enteritidis-Clade B and Enteritidis-Clade A/C are identical. SARB17 is not related to Enteritidis and is the same distance or further away than 4 other serovars.



## Appendix II: Figure S3: Phylogenetic relationship of Infantis strain SARB27.

The genome based phylogenetic tree constructed by parsnp showed that SARB27 did not cluster with either Typhimurium or Infantis. The genomes used to build this tree were from our previous study (21) except Infantis SARB27. The genome of Infantis SARB27 (RefSeq assembly accession: GCF\_000230875.1) was downloaded from NCBI (National Center for Biotechnology Information).

**SARB:** *Salmonella* Reference Collection B



28

(21): Zhang, X., Payne, M., and Lan, R. (2019). *In silico* Identification of Serovar-Specific Genes for *Salmonella* Serotyping. 10(835).  
doi:10.3389/fmicb.2019.00835.

## Appendix III: Supplementary Material of Chapter 4

The Supplementary Material for this article can be found online at:

<https://drive.google.com/drive/folders/1kEhrqzKOSWBr3ldvUvDpp9J4Y1KoMz0?usp=sharing>

### Appendix III: Data S1 Additional scripts information

Usage and further details can be found in the scripts folder at [github.com/LanLab/ShigEiFinder](https://github.com/LanLab/ShigEiFinder).

#### *clade\_specific\_gene\_combinations.py* :

This script was used to identify specific gene sets for each cluster from the pan genomes of the identification dataset. The script ran on one cluster at a time. The script takes in 4 inputs, a roary presence absence file, a genome cluster assignment file, the genomes of all isolates, the annotated genes in all genomes (as used in roary). The script first identified individual candidate genes that were present in all isolates of the target cluster (true positives) and were present only in a percentage of non-target cluster isolates (false positives). For the list of candidates each combination of genes was tested to see whether all are found in the same false positive strain. If a set of genes are never all found together then that set of genes is reported as a result. The size of the gene combinations starts at 1 for the whole list and increases progressively. At each size, successful sets of genes were reported until the total number of reported sets equals the maximum specified in the settings. Additionally, if a successful set of 2 genes (for example) was found within a subsequent set of 3 genes that three gene set was excluded because the additional gene provides no benefit.

#### *extract\_gene\_sequences\_from\_roary.py*:

This script extracts specific gene set sequences for sets produced by *clade\_specific\_gene\_combinations.py*. The script accepted 4 inputs: the presence absence roary output csv, the annotated genes in all genomes (as used in roary), a list of cluster specific genes sets and their corresponding cluster, a list of genome ids and their corresponding cluster. An output prefix is also required. The script will:

- select a representative genome from each cluster
- identify the roary orthologue group that contains a given specific gene

- retrieve the gene ID for that orthologue group and the representative genome
- extract the gene from the genes fasta file for that genome
- save the specific gene to an output file (output prefix)
- produce a summary file of genes retrieved (output prefix)

### **The selection of cluster/lineage-specific gene markers after initial screening**

- Obtain the list of genes for each set (Specific\_genes\_groupID.txt) from the output file after running script: `clade_specific_gene_combinations-fnfp.py`.
- Extract the sequences of genes using script: `prokka_genome_gene_from_roary.py`.
- Run `blastn` against identification dataset with the sequence identity of 80% to check for truncated orthologues which are not evaluated in roary.
- Gene length filtering for `blastn` output:  $\geq 50\%$  length coverage.
- Check the number of FN and FP for each cluster/lineage-specific gene set (the output file from running `clade_specific_gene_combinations-fnfp.py`), combined with the `blastn` results, the gene set with the lowest FN and FPs was selected.

### **Appendix III: Data S2 Algorithms incorporated into the ShigEiFinder**

ShigEiFinder stands for *Shigella* EIEC Cluster Enhanced Serotype Finder and is a cluster-specific gene marker based *in silico* pipeline developed for differentiation of *Shigella* and enteroinvasive *E. coli* (EIEC) and serotyping of *Shigella* and EIEC. ShigEiFinder is available as a web tool (<https://mgtdb.unsw.edu.au/ShigEiFinder/>) and on github (<https://github.com/LanLab/ShigEiFinder>).

Note that for brevity, in all references to *Shigella* serotypes below, *Shigella sonnei*, *Shigella flexneri*, *Shigella boydii* and *Shigella dysenteriae* are abbreviated as SS, SF, SB and SD respectively and a serotype is designated with abbreviated “species” name plus the serotype number e.g. *Shigella dysenteriae* serotype 1 is abbreviated as SD1.

#### **Typing reference sequences used in ShigEiFinder**

The typing reference sequences consisted of cluster-specific gene markers and sporadic EIEC lineages specific gene markers from this study, *ipaH* gene, 38 virulence genes, *Shigella* serotype specific O antigen genes collected from ShigaTyper (2), *E. coli* O antigen genes and *fliC* genes collected from SerotypeFinder (3) and 7 House Keeping (HK) genes from the MLST (4) scheme.

The cluster-specific gene marker sets and sporadic EIEC lineages specific gene markers are listed as supplementary material with file name in Table S3. The 38 virulence genes are listed in “Analysis of the 59 sporadic EIEC isolates” section in the main text. *Shigella* and *E. coli* O and H antigen genes are listed as supplementary material with file name in Data S3.

All sequences are listed in fasta format available at <https://github.com/LanLab/ShigEiFinder>.

#### **ShigEiFinder input**

Either paired end Illumina sequencing reads or assembled genomes are acceptable.

#### **ShigEiFinder output**

ShigEiFinder output included the sample, presence of *ipaH* gene, number of virulence genes, cluster assignment, serotype, *E. coli* O and H antigen present and any further notes for the result in a tabular format.

### Runtime and memory requirements

The average run time is approximately 0.89s per genome in which the average size of a genome was approximately 4.4 MB on a machine with 4 threads and 32Gb RAM.

Average script runtime for WGS reads is approximately 1.5 minutes on a machine with 4 threads and 32 Gb RAM.

### Determination of presence or absence of genes

The presence or absence of genes were determined by the cutoff value of gene length coverage for assembled genomes and the mapping length percentage and the ratio of mean mapping depth to the average mean mapping depth of 7 HK genes (Table 1). For example, the *ipaH* gene was defined as present if mapping length coverage was over 10% together with the ratio of mean depth to the average mean depth of 7 HK genes was over 1% from reads mapping.

**Table 1: Thresholds used for determination of genes present or absent**

Typing reference genes	Genomes	Reads mapping	
	Gene length coverage	Mapping length coverage	Ratio to 7 HK
<i>ipaH</i> gene	10%	10%	1%
Virulence genes	50%	50%	10%
Cluster-specific gene markers	50%	50%	10%
O antigen and H antigen genes	50%	50%	10%

### Algorithms for cluster assignment and serotyping

The *Shigella* or EIEC cluster assignment was determined by the presence of cluster-specific gene marker set that was only found within a single *Shigella* or EIEC cluster. Where marker set was used to identify a cluster, all genes must be present for a cluster to be called. ShigEiFinder also used 38 virulence genes from the pINV invasive plasmid to determine whether the plasmid was present in the isolate. When more than 25 of these genes were present, the isolate was considered to be pINV positive.

The presence of cluster-specific gene marker sets combined with the presence of *ipaH* gene and/or virulence genes the isolate was assigned to *Shigella* or EIEC cluster (Table 2).

The isolate assigned as *Shigella* or EIEC unclustered could be any new cluster that cannot be detected by any of cluster-specific gene marker set. Unclustered *Shigella* or EIEC isolate could also be those that all genes in the markers set were present but one or more of the genes from the markers set have mapping ratio between 1% and 10% and do not meet the cutoff for presence and therefore are classified as unclustered (11 isolates of 15,501 isolates in validation dataset were in that category).

**Table 2: The cluster-specific gene markers based cluster assignment**

Cluster assignment	<i>ipaH</i> gene	$\geq 26$ virulence genes	cluster-specific gene/set
<i>Shigella</i> or EIEC clusters	+	+/-	+
<i>Shigella</i> or EIEC unclustered	+	+/-	-
SB13 or SB13-atypical	-	-	+
Not <i>Shigella</i> /EIEC	-	-	-

“+”: gene presence; “+/-”: can be present or absent; “-”: gene absence.

The serotype is then assigned based on the presence of *Shigella* serotype specific O antigen genes and modification genes or *E. coli* O and H antigen genes. A “novel serotype” is assigned if there is no match to known serotypes.

#### **Low level contamination check and notes for unclustered *Shigella* or EIEC isolates**

The gene markers with mapping ratio between 1% and 10% demonstrated that the genes in the genomes may not be sequenced very well or a potential contamination. In such cases ShigEiFinder will write out a note “Possible contamination by *Shigella* or EIEC strain or low cluster-specific gene mapping depth to HK genes in cluster [cluster name]”.

The genes may have mapping ratio between 1% and 10% are listed in Table 3.

**Table 3: Gene markers with mapping ratio between 1% and 10%**



Gene markers	Number of isolates of 15,501 isolates
C1_gene_2	8
C1_gene_4	1
C5_gene_1	1
CSS_gene_3	1

#### **Additional subsets of gene markers used for *Shigella* or EIEC clusters assignment**

To increase the accuracy of typing, we added additional subsets of genes to eliminate the known false presences for cluster-specific gene markers. For example, the combination of C1 specific markers set and CSB12 specific gene marker can distinguish CSB12 from C1, if both cluster specific genes are present, the isolate is assigned CSB12 while if CSB12 specific gene is absent, the isolate is assigned as C1. There are 6 subsets of combined genes incorporated into the ShigEiFinder for elimination of false cluster assignment (Table 4).

**Table 4: Subsets of combined gene markers for elimination of false cluster assignment**

Subset 1	C1 markers set	CSB12 gene	Cluster Assignment
Isolate	+	+	CSB12
Isolate	+	-	C1
Isolate	-	+	CSB12
Subset 2	C1 markers set	CSD1 markers set	Cluster Assignment
Isolate	+	+	CSD1
Isolate	+	-	C1
Isolate	-	+	CSD1
Subset 3	C1 markers set	C2 markers set	Cluster Assignment
Isolate	+	+	C2
Isolate	+	-	C1
Isolate	-	+	C2
Subset 4	C3 markers set	C5 markers set	Cluster Assignment
Isolate	+	+	C3
Isolate	+	-	C3
Isolate	-	+	C5
Subset 5	C5 markers set	C8 markers set	Cluster Assignment
Isolate	+	+	C8
Isolate	+	-	C5
Isolate	-	+	C8
Subset 6	C2 markers set	CSS markers set	Cluster Assignment
Isolate	+	+	C2
Isolate	+	-	C2
Isolate	-	+	CSS

“+”: gene presence; “-”: gene absence.

### Serotyping SB1 and SB20 within C1

SB1 and SB20 share identical O antigen genes. For better differentiation of SB1 from SB20, we analysed C1 subbranch on the identification tree (Fig.1 in main text). The 21 isolates with presence of SB1 wzx and wzy genes were grouped into one subbranch which consisted of 2 lineages, lineage I and lineage II as Table 5.

**Table 5: The distribution of SB1 and SB20 isolates in two lineages**

Lineages	ShigaTyper assignment				
	SB1	SB20	EIEC	Untypeable	Total
Lineage I	11	0	1	2 <sup>a</sup>	14
Lineage II	4	2	2	1	9 <sup>b</sup>

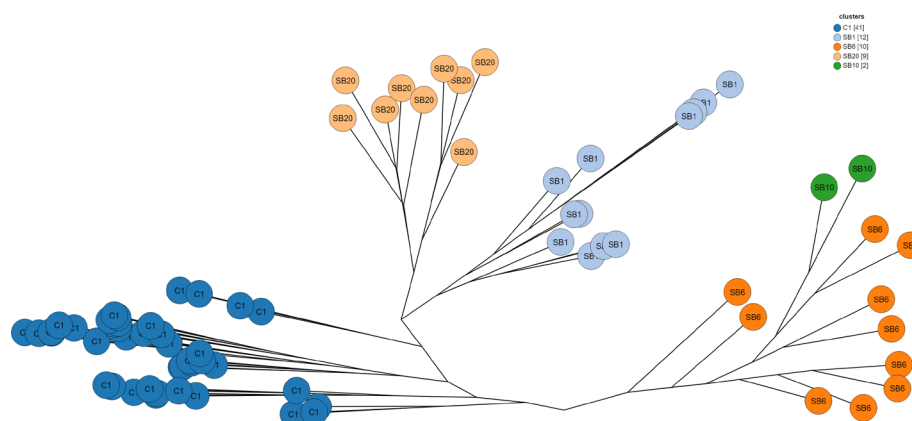
<sup>a</sup>: One isolate with the presence of heparinase gene which was used in ShigaTyper to separate SB20 from SB1. <sup>b</sup>: All 9 isolates had heparinase gene either full length or fragments by BLASTN search.

HierBAPS (5) analysis was further performed to confirm the 2 lineages. Lineage I was defined as potential SB1 lineage and lineage II was defined as SB20 lineage (Figure below). Based on phylogenetic analysis, we identified an SB20 specific gene by comparing 288 accessory genomes in C1 from the identification dataset. The gene was validated with *Shigella* and EIEC validation dataset C1 isolates. The isolate was assigned as SB20 with the presence of SB20 specific gene and SB1 wzx/wzy genes, otherwise the isolate was SB1 with the only SB1 wzx/wzy genes present.

### Serotyping SB6 and SB10 within C1

SB6 and SB10 share identical O antigen genes but there are SNP differences in the O antigen gene clusters. The SNP in SB10 wzx and SB10 wzy genes at positions 904 and 141 respectively were used to separate SB6 from SB10. For assembled genomes, we first checked the SNP positions that were covered in the blast search with 100% identify for SB10. The isolate was classified as SB10 if the SB10 SNPs were present. Otherwise, the isolate was assigned as SB6. Samtools mpileups was used to gather the nucleotide base at the SNP positions for reads mapping. The isolate was SB10 if the SB10-SNPs was found. The absence of the SNP was assigned as SB6.

### Figure: Subbranch of C1 on identification tree



### Serotyping EIEC O164/O124

The *E. coli* O164 and O124 O antigen genes are near identical with > 99.4% identity (6). There was a 2-base indel (a frame shift mutation (7)) at positions 429 and 430 in *wfeP* gene of O164 in comparison to O124. We used this indel to differentiate O164 from O124. The isolate was assigned as O164 if the indel was found.

### Multiple variants of H antigens

There are multiple variants for one type of H antigen. To assign an H type when multiple H variants are present, the highest match was chosen as the H antigen present.

### SF serotyping within C3

C3 contains all SF serotypes except for SF6 which is grouped into C1. We used the established scheme of SF O antigen genes and modification genes including *gtr*, *oac* and *opt* genes to type SF within C3 (8-20) (Table 6). ShigEiFinder assigned all possibilities when there was a multiple match of combinations of modification genes. The isolate was classified as SFY if there was only backbone O antigen genes present. While the isolate was assigned as SF novel serotype if no match to known serotypes and the note was given with the presence or absence of genes.

**Table 6: The combination of O antigen genes and modification genes used for SF serotyping**

	wzx <sub>1-5</sub>	wzx <sub>6</sub>	gtrI	gtrIC	gtrII	gtrIV	gtrV	gtrX	oac	oacIb	oacB	oacC	oacD	optII	optIII
SF1a	+	-	+	-	-	-	-	-	-	-	+/-	-	-	-	-
SF1b	+	-	+	-	-	-	-	-	-	+	+/-	-	-	-	-
SF1c(7a)	+	-	+	+	-	-	-	-	-	-	-	-	-	-	-
SF1d	+	-	+	-	-	-	-	+	-	-	-	-	-	-	-
SF2a	+	-	-	-	+	-	-	-	-	-	+/-	-	+/-	-	-
SF2b	+	-	-	-	+	-	-	+	-	-	-	-	+/-	-	-
SF3a	+	-	-	-	-	-	-	+	+/-	-	-	-	+/-	-	-
SF3b	+	-	-	-	-	-	-	-	+	-	-	-	-	-	-
SF4a	+	-	-	-	-	+	-	-	-	-	-	-	-	-	-
SF4av	+	-	-	-	-	+	-	-	-	-	-	-	-	-	+
SF4b	+	-	-	-	-	+	-	-	+	-	-	-	-	-	-
SF5a	+	-	-	-	-	-	+	-	-	-	+	-	-	-	-
SF5b	+	-	-	-	-	-	+	+	-	-	-	-	-	-	-
SF7b	+	-	-	+	-	-	-	-	+	-	-	-	-	-	-
SFX	+	-	-	-	-	-	-	+	-	-	-	-	+/-	-	-
SFXv(4c)	+	-	-	-	-	-	-	+	-	-	-	-	+/-	+	-
SFY	+	-	-	-	-	-	-	-	-	-	+/-	-	+/-	-	-
SFYv	+	-	-	-	-	-	-	-	-	-	-	-	+	+	+
SF6	-	+	-	-	-	-	-	-	-	-	-	+/-	-	-	-

“+”: gene presence and highlighted in pink color. “+/-”: can be present or absent. “-”: gene absence.

#### Reference:

1. Zhang X, Payne M, Lan R. *In silico* Identification of Serovar-Specific Genes for *Salmonella* Serotyping. *Frontiers in microbiology*. 2019;10:835.
2. Wu Y, Lau HK, Lee T, Lau DK, Payne J. *in Silico* Serotyping Based on Whole-Genome Sequencing Improves the Accuracy of *Shigella* Identification. *Applied and environmental microbiology*. 2019;85(7).
3. Joensen KG, Tetzschner AM, Iguchi A, Aarestrup FM, Scheutz F. Rapid and Easy *In Silico* Serotyping of *Escherichia coli* Isolates by Use of Whole-Genome Sequencing Data. *Journal of clinical microbiology*. 2015;53(8):2410-26.
4. Jolley KA, Maiden MC. BIGSdb: Scalable analysis of bacterial genome variation at the population level. *BMC Bioinformatics*. 2010;11:595.

5. Cheng L, Connor TR, Sirén J, Aanensen DM, Corander J. Hierarchical and spatially explicit clustering of DNA sequences with BAPS software. *Molecular biology and evolution*. 2013;30(5):1224-8.
6. Liu B, Furevi A, Perepelov AV, Guo X, Cao H, Wang Q, et al. Structure and genetics of *Escherichia coli* O antigens. *FEMS microbiology reviews*. 2020;44(6):655-83.
7. Liu B, Knirel YA, Feng L, Perepelov AV, Senchenkova SN, Wang Q, et al. Structure and genetics of *Shigella* O antigens. *FEMS microbiology reviews*. 2008;32(4):627-53.
8. Adhikari P, Allison G, Whittle B, Verma NK. Serotype 1a O-antigen modification: molecular characterization of the genes involved and their novel organization in the *Shigella flexneri* chromosome. *Journal of bacteriology*. 1999;181(15):4711-8.
9. Stagg RM, Tang SS, Carlin NI, Talukder KA, Cam PD, Verma NK. A novel glucosyltransferase involved in O-antigen modification of *Shigella flexneri* serotype 1c. *Journal of bacteriology*. 2009;191(21):6612-7.
10. Mavris M, Manning PA, Morona R. Mechanism of bacteriophage SfII-mediated serotype conversion in *Shigella flexneri*. *Molecular microbiology*. 1997;26(5):939-50.
11. Adams MM, Allison GE, Verma NK. Type IV O antigen modification genes in the genome of *Shigella flexneri* NCTC 8296. *Microbiology (Reading, England)*. 2001;147(Pt 4):851-60.
12. Huan PT, Bastin DA, Whittle BL, Lindberg AA, Verma NK. Molecular characterization of the genes involved in O-antigen modification, attachment, integration and excision in *Shigella flexneri* bacteriophage SfV. *Gene*. 1997;195(2):217-27.
13. Verma NK, Verma DJ, Huan PT, Lindberg AA. Cloning and sequencing of the glucosyl transferase-encoding gene from converting bacteriophage X (SFX) of *Shigella flexneri*. *Gene*. 1993;129(1):99-101.
14. Clark CA, Beltrame J, Manning PA. The *oac* gene encoding a lipopolysaccharide O-antigen acetylase maps adjacent to the integrase-encoding gene on the genome of *Shigella flexneri* bacteriophage Sf6. *Gene*. 1991;107(1):43-52.
15. Sun Q, Lan R, Wang Y, Wang J, Xia S, Wang Y, et al. Identification of a divergent O-acetyltransferase gene *oac* 1b from *Shigella flexneri* serotype 1b strains. *Emerging microbes & infections*. 2012;1(9):e21.
16. Wang J, Knirel YA, Lan R, Senchenkova SN, Luo X, Perepelov AV, et al. Identification of an O-acyltransferase gene (*oacB*) that mediates 3- and 4-O-acetylation

of rhamnose III in *Shigella flexneri* O antigens. *Journal of bacteriology*. 2014;196(8):1525-31.

17. Knirel YA, Wang J, Luo X, Senchenkova SN, Lan R, Shpirt AM, et al. Genetic and structural identification of an O-acyltransferase gene (*oacC*) responsible for the 3/4-O-acetylation on rhamnose III in *Shigella flexneri* serotype 6. *BMC microbiology*. 2014;14:266.

18. Sun Q, Knirel YA, Wang J, Luo X, Senchenkova SN, Lan R, et al. Serotype-converting bacteriophage SfII encodes an acyltransferase protein that mediates 6-O-acetylation of GlcNAc in *Shigella flexneri* O-antigens, conferring on the host a novel O-antigen epitope. *Journal of bacteriology*. 2014;196(20):3656-66.

19. Sun Q, Knirel YA, Lan R, Wang J, Senchenkova SN, Jin D, et al. A novel plasmid-encoded serotype conversion mechanism through addition of phosphoethanolamine to the O-antigen of *Shigella flexneri*. *PloS one*. 2012;7(9):e46095.

20. Knirel YA, Lan R, Senchenkova SN, Wang J, Shashkov AS, Wang Y, et al. O-antigen structure of *Shigella flexneri* serotype Yv and effect of the *lpt-O* gene variation on phosphoethanolamine modification of *S. flexneri* O-antigens. *Glycobiology*. 2013;23(4):475-85.

### Appendix III: Data S3 *Shigella*/EIEC serotypes specific O and H antigens used in ShigEiFinder

*Shigella* serotype specific O antigen genes were collected from ShigaTyper (2). *E. coli* O antigen genes and *fliC* genes were collected from SerotypeFinder (3)

Sequences	Accession number
SD1_wzx, SD1_wzy	L07293
SD1-rfp	CP000640
SD2_wzx, SD2_wzy	EU296404
SD3_wzx, SD3_wzy	EU296415
SD4_wzx, SD4_wzy	EU296402
SD5_wzx, SD5_wzy	EU294174
SD6_wzx, SD6_wzy	EU296414
SD7_wzx, SD7_wzy	AY380835
SD8_wzx, SD8_wzy	EU294166
SD9_wzx, SD9_wzy	EU296416
SD10_wzx, SD10_wzy	EU294178
SD11_wzx, SD11_wzy	EU294172
SD12_wzx, SD12_wzy	EU294169
SD13_wzx, SD13_wzy	EU294167
SD14_wzx, SD14_wzy	CP026832
SD15_wzx, SD15_wzy	CP026834
SDP 96-265_wzx, wzy	CP026819
SDP E670-74_wzx,wzy	CP027027
SB1_wzx, SB1_wzy	AY630255
SB2_wzx, SB2_wzy	EU296418
SB3_wzx, SB3_wzy	EU296407
SB4_wzx, SB4_wzy	AF402312
SB5_wzx, SB5_wzy	AF402313
SB6_wzx, SB6_wzy	AF402314
SB7_wzx, SB7_wzy	EU296411
SB8_wzx, SB8_wzy	EU294163
SB9_wzx, SB9_wzy	AF402315
SB10_wzx, SB10_wzy	AY693427
WbaM	AY693427
SB11_wzx, SB11_wzy	AY529126
SB12_wzx, SB12_wzy	EU296406
SB13_wzx, SB13_wzy	AY369140
SB14_wzx, SB14_wzy	EU296409
SB15_wzx, SB15_wzy	EU296412
SB16_wzx, SB16_wzy	DQ371800
SB17_wzx, SB17_wzy	DQ875941
SB18_wzx, SB18_wzy	AY948196
SB19_wzx, SB19_wzy	CP026814

Heparinase	CP016036
SBP E1621-54_wzx,wzy	CP026810
SF <i>wzx</i> <sub>1-5</sub> gene	AE005674
SF6 <i>wzx</i> gene	EU294165
SF <i>gtrI</i>	AF139596
SF <i>gtrIC</i>	FJ905303
SF <i>gtrII</i>	AF021347
SF <i>gtrIV</i>	AF288197
SF <i>gtrV</i>	U82619
SF <i>gtrX</i>	L05001
SF <i>oacA</i>	AF547987
SF <i>oacIb</i>	JN377795
SF <i>oacB</i>	NC_004337 (SF0315)
SF <i>oacC</i>	AKMW01000058
SF <i>oacD</i>	NC_004337 (SF0309)
SF <i>optIII</i>	KC020049
SF Xv <i>optII</i>	CP001385 (SFxv_5135)
SS_wzx, SS_wzy	AF285971
O1_wzx, O1_wzy	GU299791
O2_wzx, O2_wzy	EU549863
O4_wzx, O4_wzy	AY568960
O6_wzx, O6_wzy	AJ426045
O7_wzx, O7_wzy	AF125322
O8_wzx, O8_wzy	AF013583
O8_wzm, O8_wzt	AB010150
O12_wzx, O12_wzy	AB811600
O13_wzx, O13_wzy	EU296422
O16_wzx, O16_wzy	AB811601
O17_wzx, O17_wzy	AB812084
O18ac_wzx, O18ac_wzy	AB811603
O21_wzx, O21_wzy	EU694098
O22_wzx, O22_wzy	AB811606
O25_wzx, O25_wzy	GU014554
O26_wzx, O26_wzy	AF529080
O28ac_wzx, O28ac_wzy	DQ462205
O29_wzx, O29_wzy	EU294173
O32_wzx, O32_wzy	EU296410
O36_wzx, O36_wzy	AB811613
O39_wzx, O39_wzy	AB811616
O40_wzx, O40_wzy	EU296417
O50_wzx, O50_wzy	AB811624
O53_wzx, O53_wzy	EU289392
O71_wzx, O71_wzy	GU445927
O77_wzx, O77_wzy	AB972416
O79_wzx, O79_wzy	EU294162
O86_wzx, O86_wzy	AY220982



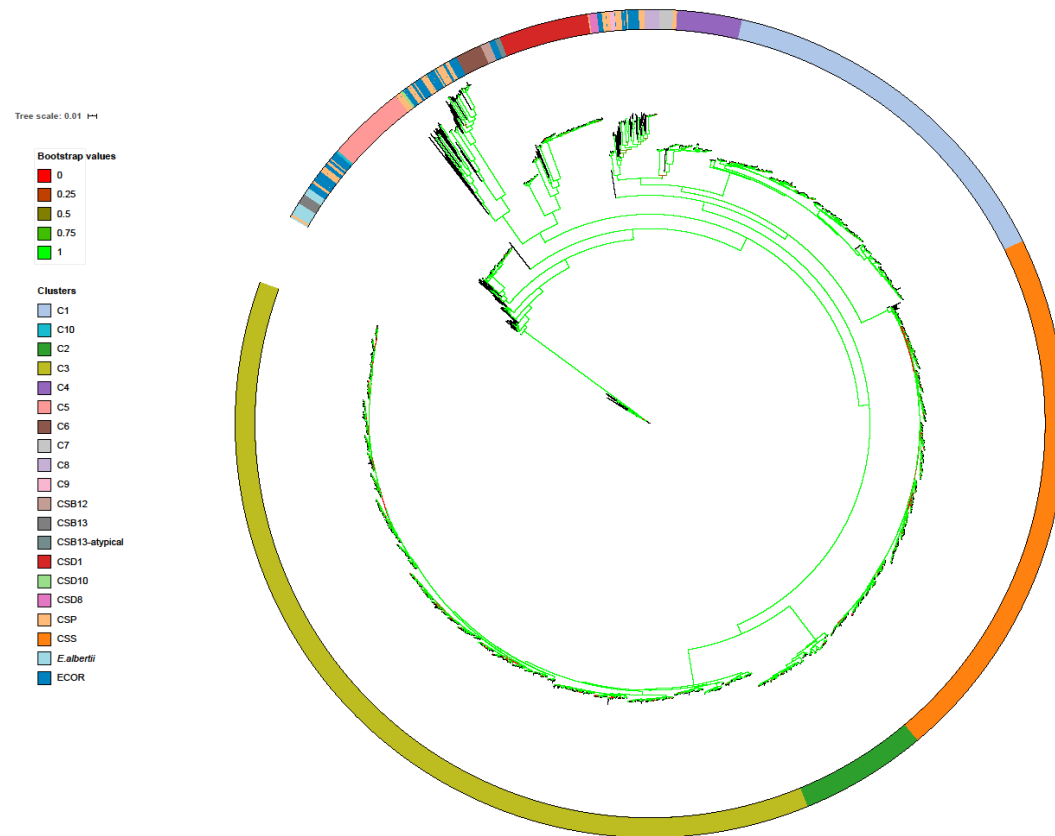
O89_wzx, O89_wzy	AB812038
O92_wzx, O92_wzy	AB812040
O93_wzx, O93_wzy	AB812041
O96_wzx, O96_wzy	AB812043
O102_wzx, O102_wzy	JX087966
O105_wzx, O105_wzy	EU294171
O110_wzx, O110_wzy	AB812049
O111_wzx, O111_wzy	JN887675
O112ab_wzx, O112ab_wzy	EU296413
O112ac_wzx, O112ac_wzy	EU296405
O117_wzx, O117_wzy	EU694096
O118_wzx, O118_wzy	HM204927
O121_wzx, O121_wzy	JN859209
O124_wzx, O124_wzy	EU296419
O129_wzx, O129_wzy	EU296424
O130_wzx, O130_wzy	EU296421
O132_wzx, O132_wzy	AB812056
O135_wzx, O135_wzy	EU296423
O136_wzx, O136_wzy	AB812059
O143_wzx, O143_wzy	EU294164
O144_wzx, O144_wzy	AB812062
O147_wzx, O147_wzy	DQ868766
O148_wzx, O148_wzy	DQ167407
O149_wzx, O149_wzy	DQ868764
O151_wzx, O151_wzy	HM204926
O152_wzx, O152_wzy	EU294170
O155_wzx, O155_wzy	AY657020
O162_wzx, O162_wzy	AB812067
O164_wzx, O164_wzy	EU296420
O167_wzx, O167_wzy	EU296408
O173_wzx, O173_wzy	GU068046
O180_wzx, O180_wzy	JQ751058
O183_wzx, O183_wzy	AB627352
H1_fliC	AB028471
H2_fliC	AIHA01000023
H4_fliC	AJ605764
H4_fliC	AJ605765
H4_fliC	AJ536600
H5_fliC	AY249990
H5_fliC	AY337469
H6_fliC	AIEY01000041
H7_fliC	AY337468
H7_fliC	AKML01000326
H7_fliC	ANLT01000257
H7_fliC	ANLJ01000383
H7_fliC	AOES01000098
H7_fliC	AMVH01000352

H7_ <i>fliC</i>	AF228487
H7_ <i>fliC</i>	AF228496
H7_ <i>fliC</i>	AF228495
H7_ <i>fliC</i>	AB334575
H7_ <i>fliC</i>	AB334574
H7_ <i>fliC</i>	AF228494
H7_ <i>fliC</i>	AF228491
H7_ <i>fliC</i>	AF228492
H7_ <i>fliC</i>	AB028474
H8_ <i>fliC</i>	AJ865465
H9_ <i>fliC</i>	AY249994
H10_ <i>fliC</i>	AF169320
H11_ <i>fliC</i>	AY337472
H12_ <i>fliC</i>	AY337471
H14_ <i>fliC</i>	AY249998
H16_ <i>fliC</i>	AB128919
H16_ <i>fliC</i>	JH954529
H16_ <i>fliC</i>	JH953794
H16_ <i>fliC</i>	AY337476
H16_ <i>fliC</i>	AY337477
H16_ <i>fliC</i>	AY337475 AY2500001
H16_ <i>fliC</i>	AY250000
H18_ <i>fliC</i>	AY250001
H19_ <i>fliC</i>	AY337479
H19_ <i>fliC</i>	AY250002
H20_ <i>fliC</i>	AY250003
H21_ <i>fliC</i>	AIHL01000060
H24_ <i>fliC</i>	K72 (H25w)
H25_ <i>fliC</i>	AGSG01000116
H26_ <i>fliC</i>	AY250008
H26_ <i>fliC</i>	AY337483
H27_ <i>fliC</i>	AY250009
H28_ <i>fliC</i>	AY250010
H30_ <i>fliC</i>	AY250011
H30_ <i>fliC</i>	AY337483
H31_ <i>fliC</i>	AY250013
H33_ <i>fliC</i>	AY250015
H40_ <i>fliC</i>	AJ884568
H42_ <i>fliC</i>	AY250021
H45_ <i>fliC</i>	AY250023
H48_ <i>fliC</i>	AY250025
H49_ <i>fliC</i>	AY250026
H51_ <i>fliC</i>	AY250027

**Reference:**

1. Wu Y, Lau HK, Lee T, Lau DK, Payne J. *in Silico* Serotyping Based on Whole-Genome Sequencing Improves the Accuracy of *Shigella* Identification. *Applied and environmental microbiology*. 2019;85(7).
2. Joensen KG, Tetzschner AM, Iguchi A, Aarestrup FM, Scheutz F. Rapid and Easy *In Silico* Serotyping of *Escherichia coli* Isolates by Use of Whole-Genome Sequencing Data. *Journal of clinical microbiology*. 2015;53(8):2410-26.

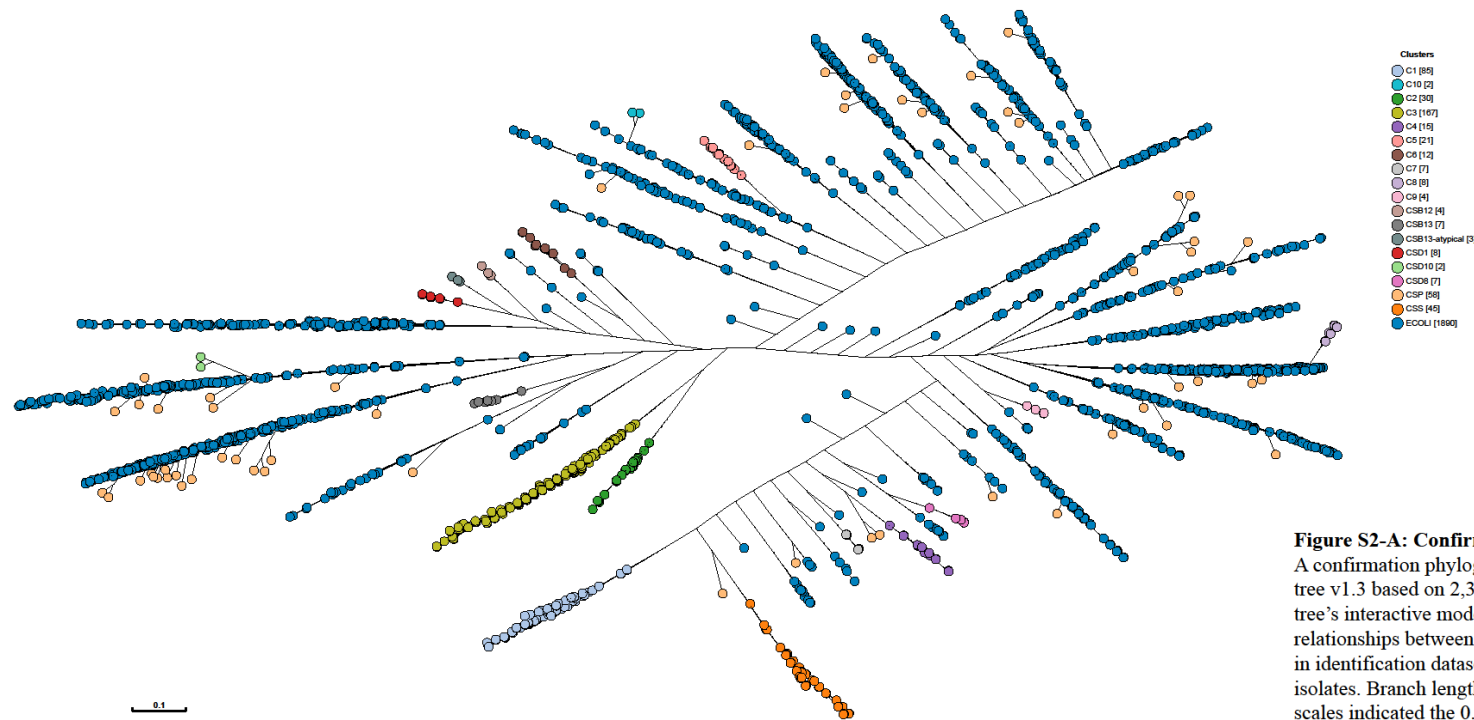
### Appendix III: Figure S1: Identification phylogenetic tree



#### Figure S1: Identification phylogenetic tree

An identification phylogenetic tree constructed by Quicktree v1.3 and visualised by ITOL v5 shows the phylogenetic relationships of 1,879 *Shigella* and EIEC isolates in identification dataset. The tree scales indicated the 0.01 substitutions per locus. *Shigella* and EIEC clusters are colored. The internal branches are colored to represent the bootstrap values. Green color indicates the maximum bootstrap value (1). The red color shows the minimum bootstrap value (0). Each of cluster is well supported by bootstrap value of 80% or greater. CSP is sporadic EIEC lineages. ECOR is *Escherichia coli* reference collection.

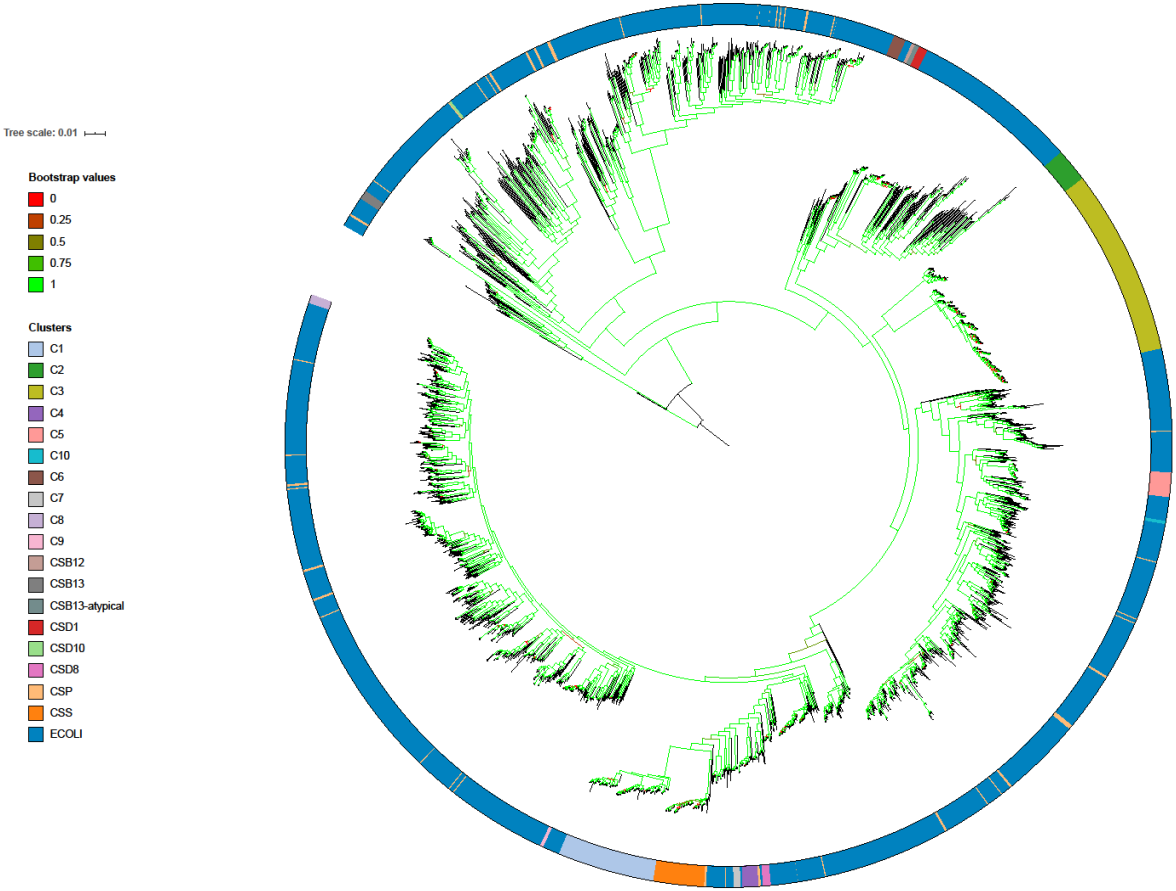
### Appendix III: Figure S2-A: Confirmation phylogenetic tree



**Figure S2-A: Confirmation phylogenetic tree**

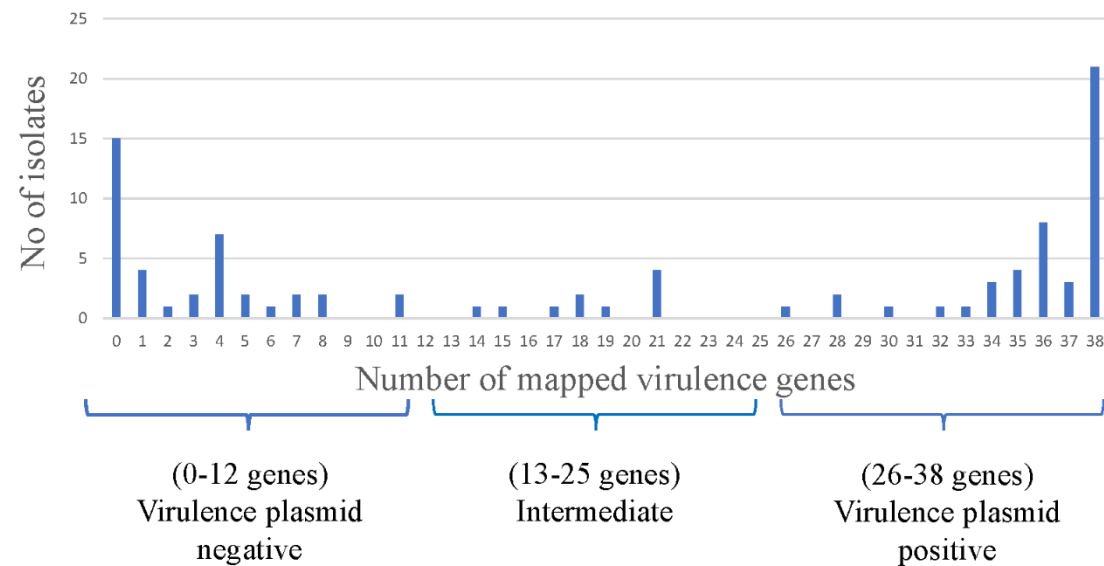
A confirmation phylogenetic tree was constructed by Quick-tree v1.3 based on 2,375 isolates and visualised by Grape-tree's interactive mode. The tree shows the phylogenetic relationships between identified *Shigella* and EIEC clusters in identification dataset and non-enteroinvasive *E. coli* isolates. Branch lengths are log scale for clarity. The tree scales indicated the 0.1 substitutions per locus. Known *Shigella* and EIEC clusters from identification dataset are colored. Numbers in square brackets indicate the number of isolates of each identified cluster. CSP is sporadic EIEC lineages.

Appendix III: Figure S2-B: Confirmation phylogenetic tree



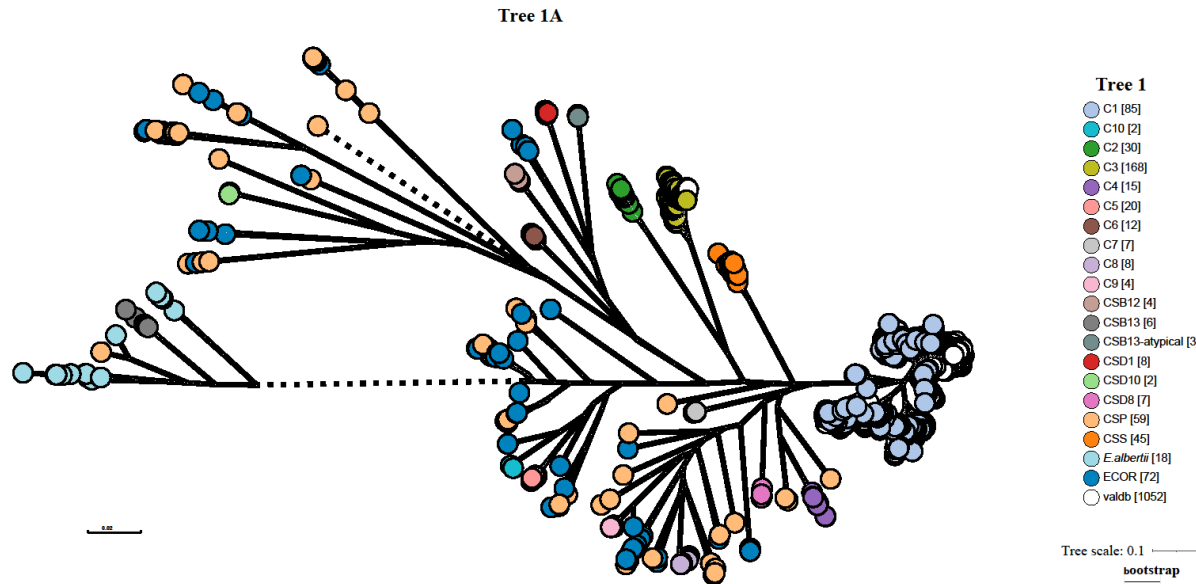
**Figure S2-B: Confirmation phylogenetic tree**  
A confirmation phylogenetic tree constructed by Quicktree v1.3 and visualised by ITOL v5 shows the phylogenetic relationships between identified *Shigella* and EIEC clusters in identification dataset and non-enteroinvasive *E. coli* isolates. The tree scales indicated the 0.01 substitutions per locus. *Shigella* and EIEC clusters are colored. The internal branches are colored to represent the bootstrap values. Green color indicates the maximum bootstrap value (1). The red color shows the minimum bootstrap value (0). Each of cluster is well supported by bootstrap value of 80% or greater. CSP is sporadic EIEC lineages.

### Distribution of mapped 38 virulence genes in 59 sporadic isolates

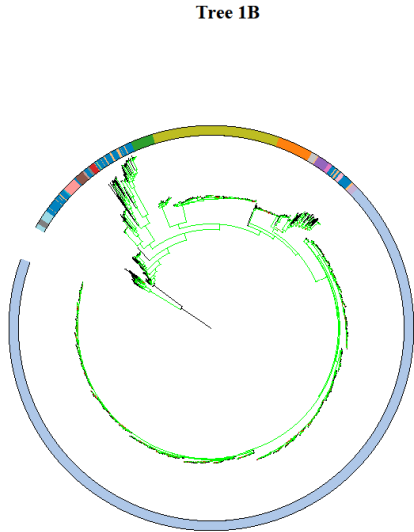


**Appendix III: Figure S3: Distribution of mapped 38 virulence genes in 58 sporadic isolates.** The presence of *Shigella* virulence plasmid pINV in 58 sporadic isolates in identification dataset was determined by the mapped 38 virulence genes. Detailed genes were described in Results “Investigation of *Shigella* virulence plasmid pINV in 59 sporadic isolates”. Three categories were defined based on the number of virulence genes mapped to isolate. Virulence plasmid positive: > 25 genes mapped to isolate; Intermediate: 13 to 25 genes mapped to isolate; Virulence plasmid negative: less than 13 genes mapped to isolate.

Appendix III: Figure S4: Validation phylogenetic trees



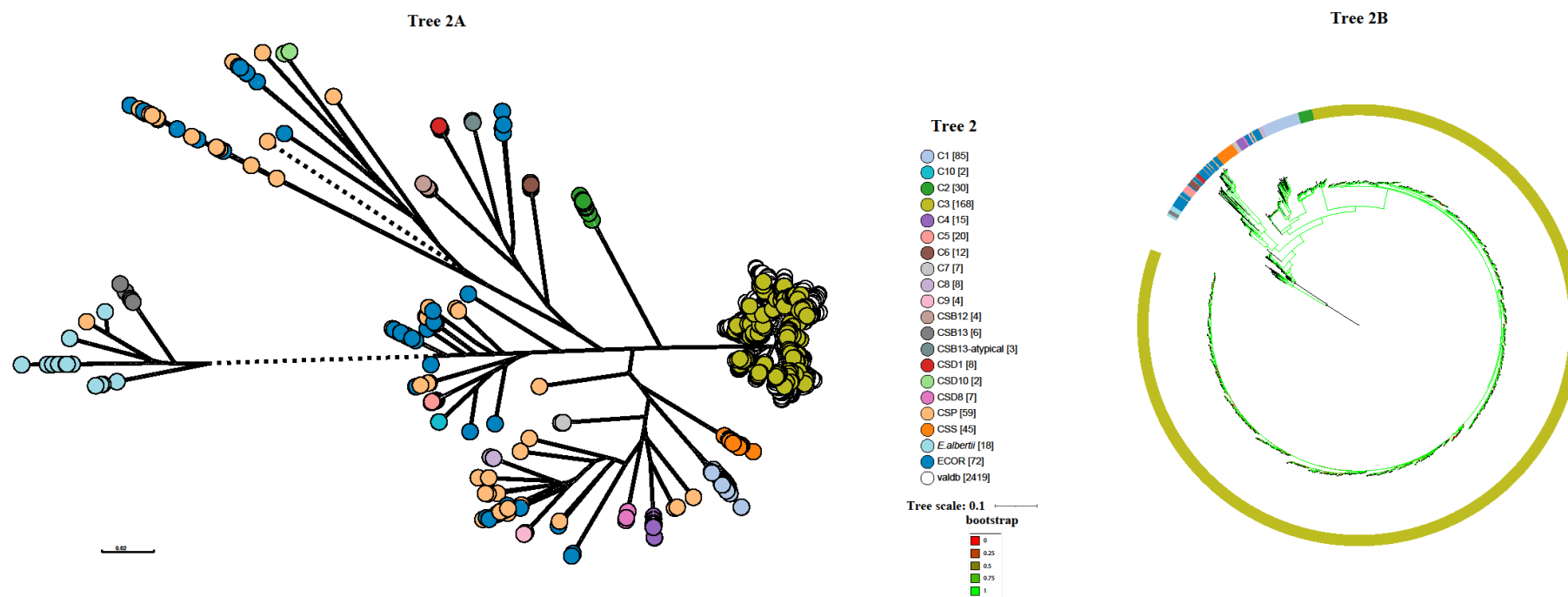
**Figure S4-A: Validation phylogenetic tree**  
Seven validation trees were generated by Quicktree v1.3 and visualised by Grapetree's to assign isolates in validation dataset to clusters. Branch lengths are log scale for clarity. The tree scales indicated the 0.2 substitutions per locus. Known *Shigella* and EIEC clusters from identification dataset are colored. Numbers in square brackets indicate the number of isolates of each identified cluster. Isolates in validation dataset (valdb) are colored white. The isolates were assigned to clusters if the isolates were found within a branch that exclusively contained identification dataset isolates from that clusters with a bootstrap value of 80% or greater. CSP is sporadic EIEC lineages.



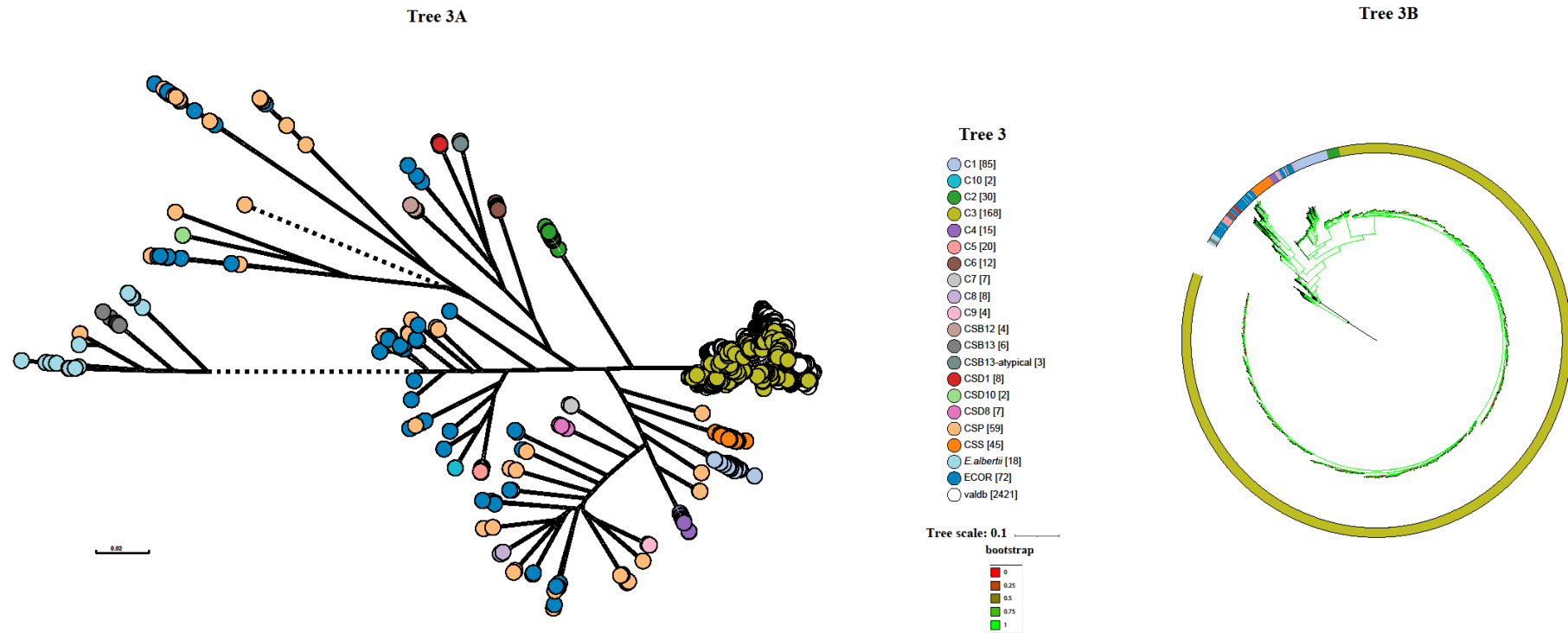
**Figure S4-B: Validation phylogenetic tree**  
Seven validation trees were constructed by Quicktree v1.3 and visualised by ITOL v5 to assign isolates in validation dataset to clusters. The tree scales indicated the 0.01 substitutions per locus. *Shigella* and EIEC clusters are colored. The internal branches are colored to represent the bootstrap values. Green color indicates the maximum bootstrap value (1). The red color shows the minimum bootstrap value (0). Each of cluster is well supported by bootstrap value of 80% or greater. Isolates that grouped with known cluster isolates (from identification dataset) with strong bootstrap support are categorized into that cluster. CSP is sporadic EIEC lineages.



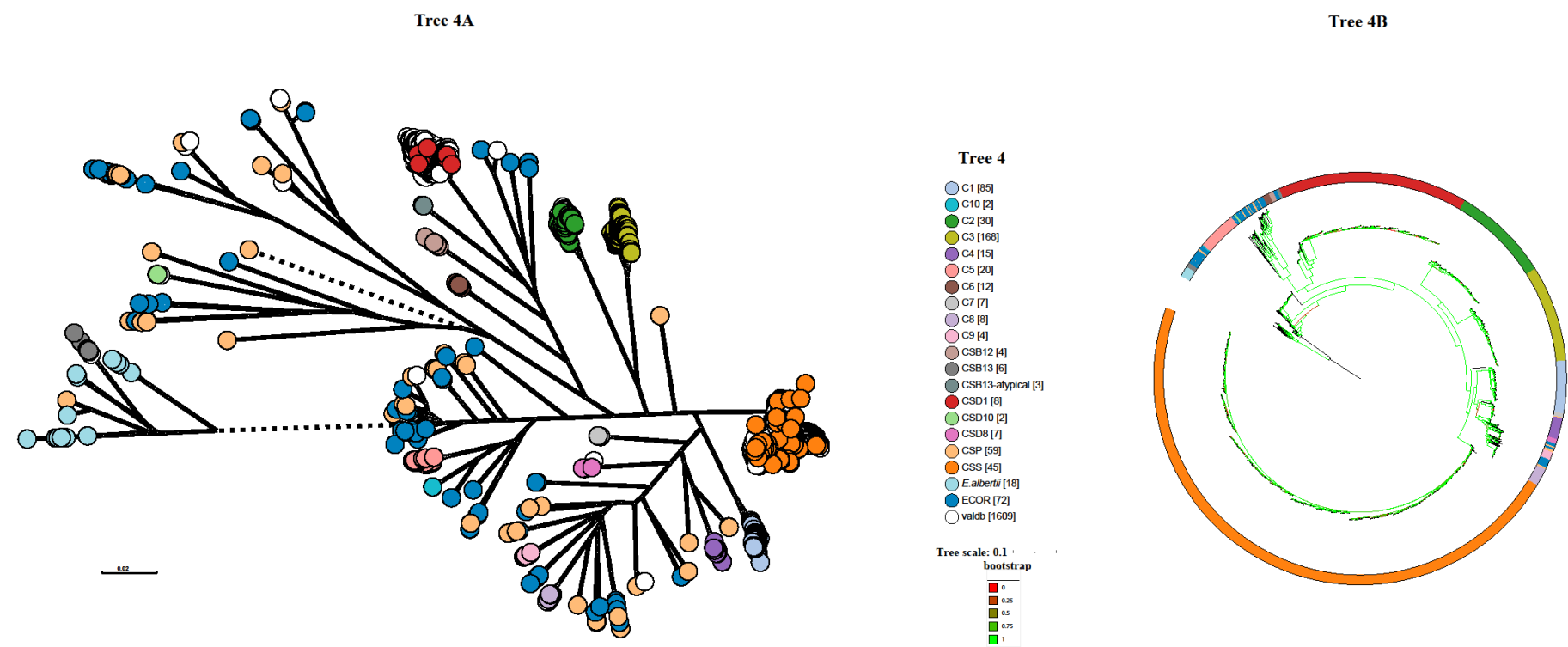
### Appendix III: Figure S4: Validation phylogenetic trees



### Appendix III: Figure S4: Validation phylogenetic trees

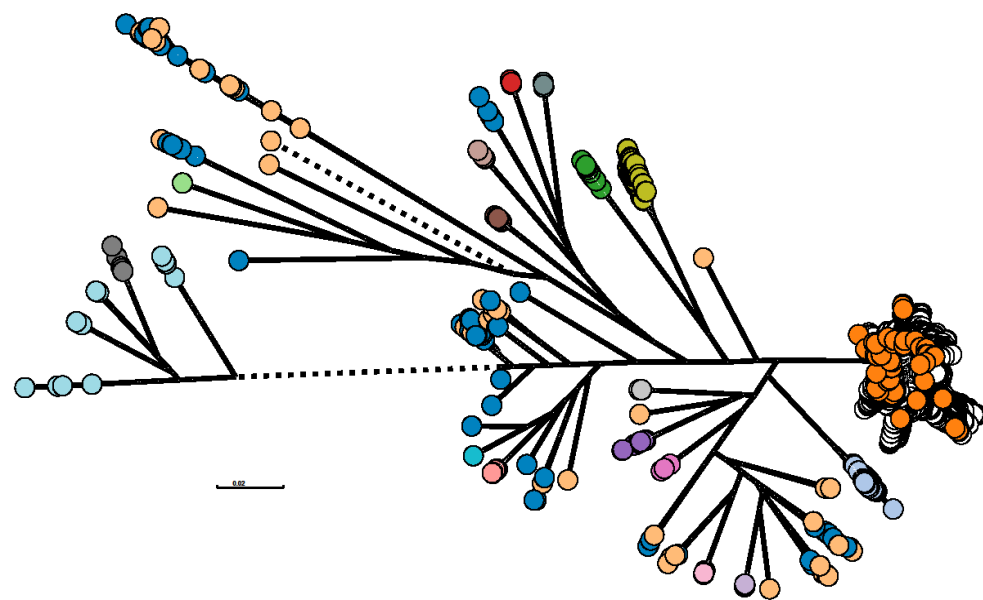


Appendix III: Figure S4: Validation phylogenetic trees

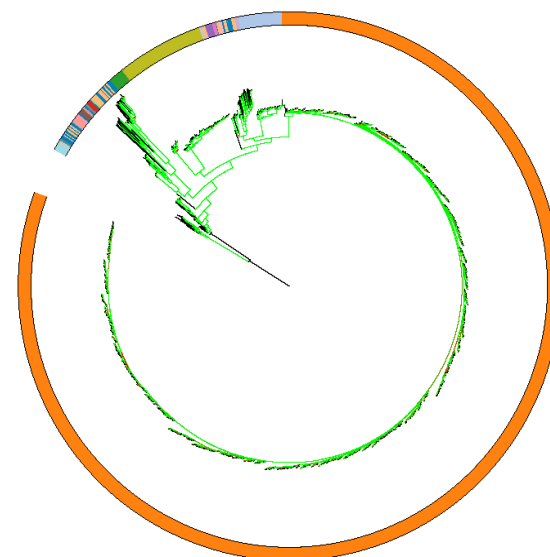


Appendix III: Figure S4: Validation phylogenetic trees

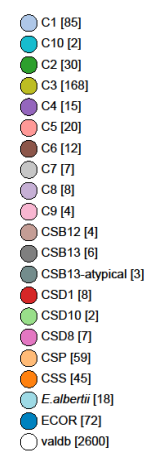
Tree 5A



Tree 5B



Tree 5

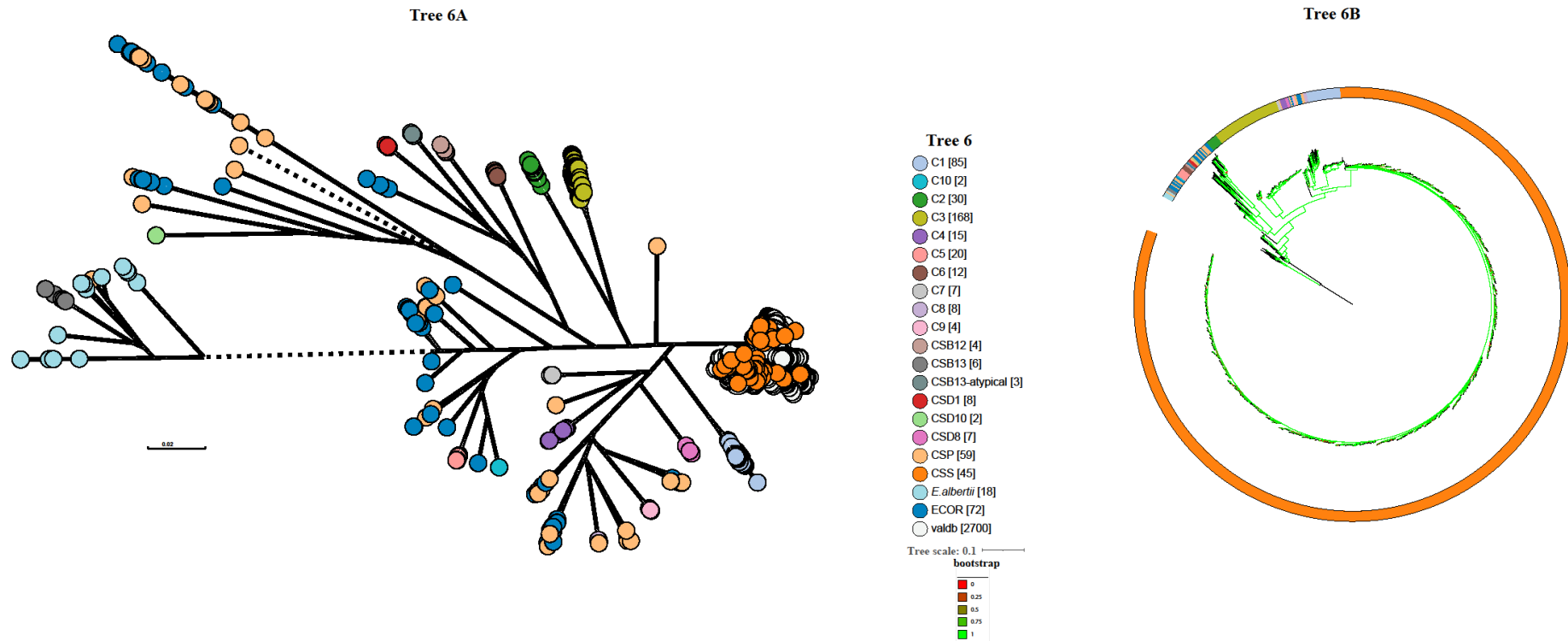


Tree scale: 0.1

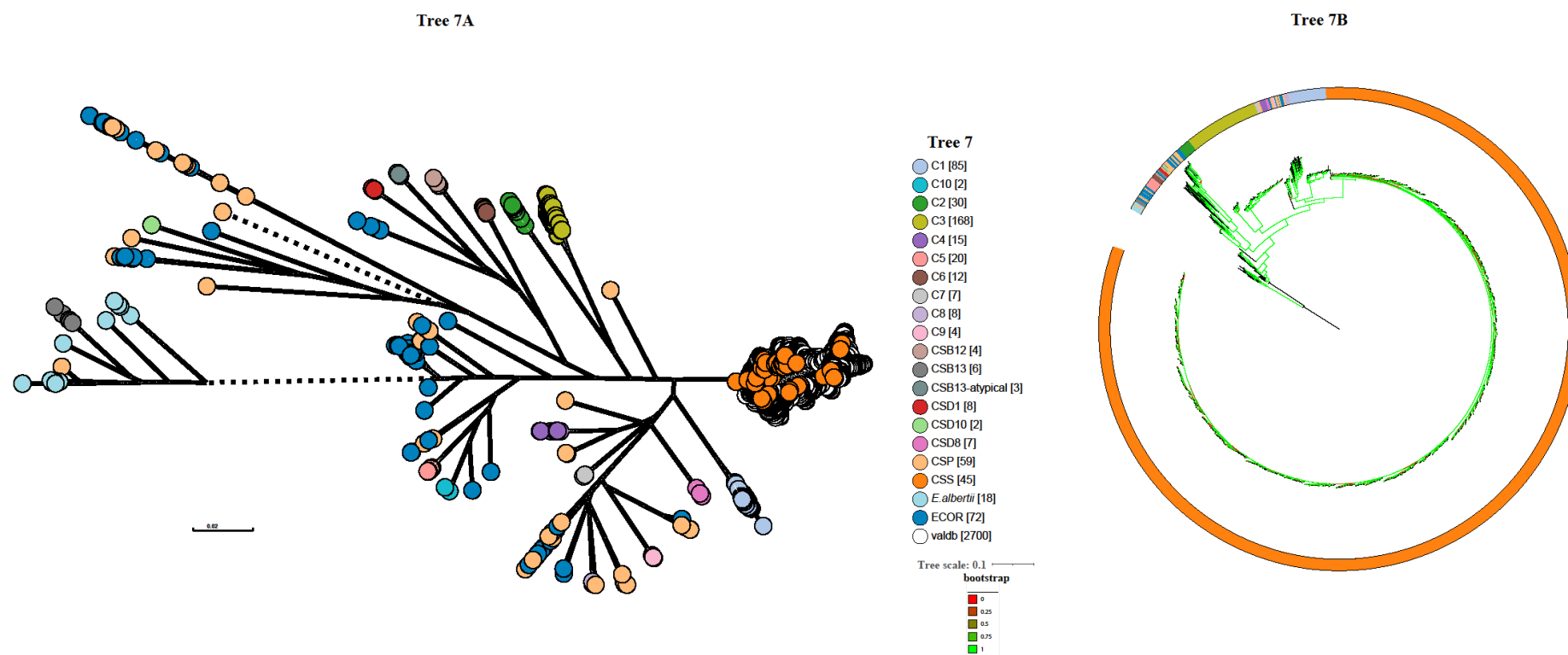
bootstrap



# Appendix III: Figure S4: Validation phylogenetic trees



### Appendix III: Figure S4: Validation phylogenetic trees



## **Appendix IV: Supplementary Material of Chapter 5**

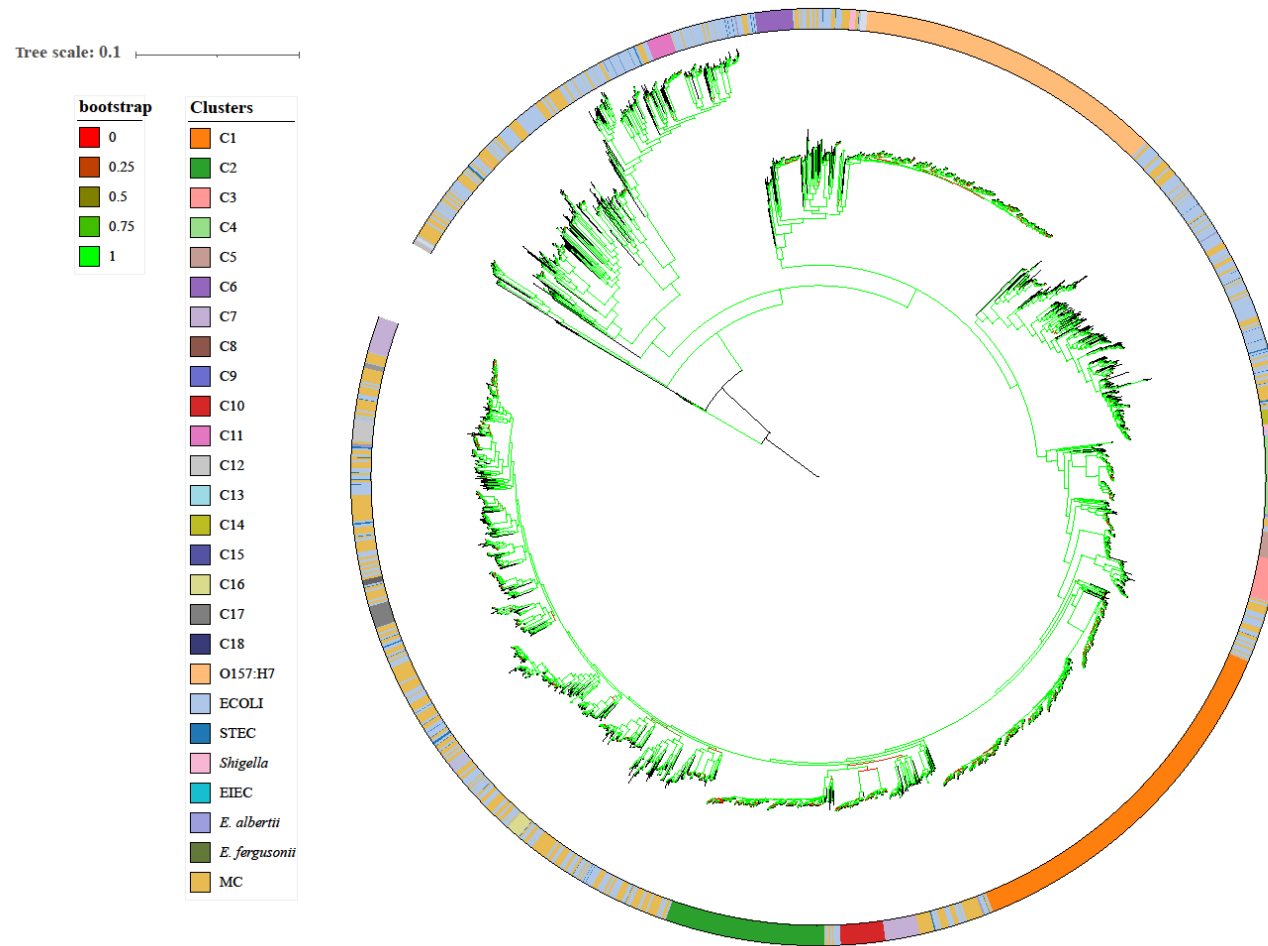
**The Supplementary Material for this article can be found online at:**

<https://drive.google.com/drive/folders/1HXJvKlHHYeQ9-ZoCQ7WnY7I3oX4lhUfW?usp=sharing>

**Appendix IV: Supplementary Tables:**

<https://drive.google.com/drive/folders/1HXJvKlHHYeQ9-ZoCQ7WnY7I3oX4lhUfW?usp=sharing>

## Appendix IV: Figure S1: Identification phylogenetic tree

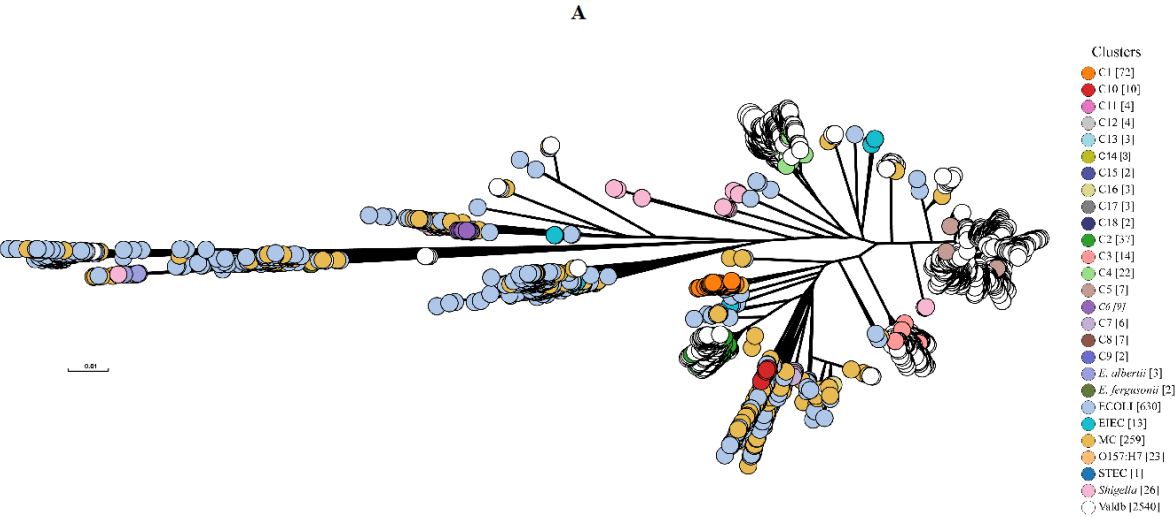


**Figure S1: Identification phylogenetic tree**  
An identification phylogenetic tree constructed by Quicktree v1.3 and visualised by ITOL v5 shows the phylogenetic relationships of 2567 STEC isolates in identification dataset. The tree scales indicated the 0.01 substitutions per locus. STEC clusters are colored. The internal branches are colored to represent the bootstrap values. Green color indicates the maximum bootstrap value (1). The red color shows the minimum bootstrap value (0). Each of cluster is well supported by bootstrap value of 80% or greater. MC is STEC minor clusters.

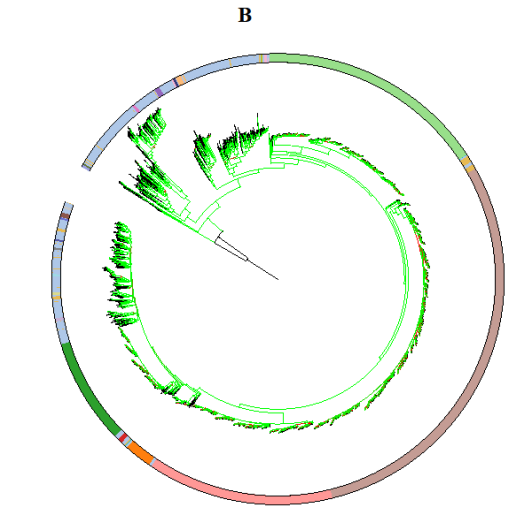


Appendix IV: Figure S2: Validation phylogenetic tree

Figure S2: The representative validation phylogenetic tree



**Figure S2-A: The representative validation phylogenetic tree**  
Fifteen validation trees were generated by Quicktree v1.3 and visualised by Grapetree to assign isolates in validation dataset to clusters. The one representative tree is shown in detail as an example and all the others are similar. The tree scales indicated the 0.01 substitutions per locus. Known STEC clusters from identification dataset are colored. Numbers in square brackets indicate the number of isolates of each identified cluster. Isolates in validation dataset (valdb) are colored in white. The isolates were assigned to clusters if the isolates were found within a branch that exclusively contained identification dataset isolates from that clusters with a bootstrap value of 80% or greater. MC is STEC minor clusters.



**Figure S2-B: The representative validation phylogenetic tree**  
Fifteen validation trees were constructed by Quicktree v1.3 and visualised by ITOL v5 to assign isolates in validation dataset to clusters. The one representative tree is shown in detail as an example and all the others are similar. The tree scales indicated the 0.01 substitutions per locus. STEC clusters are colored. The internal branches are colored to represent the bootstrap values. Green color indicates the maximum bootstrap value (1). The red color shows the minimum bootstrap value (0). Each of cluster is well supported by bootstrap value of 80% or greater. Isolates that grouped with known cluster isolates (from identification dataset) with strong bootstrap support are categorized into that cluster. MC is STEC minor clusters.

