SALMONELLA SEROTYPING AND SOURCE PREDICTION USING WHOLE GENOME SEQUENCING DATA

by

SHAOKANG ZHANG

(Under the Direction of Dr. Xiangyu Deng)

ABSTRACT

In our study, we utilized whole genome sequencing (WGS) and bioinformatics approaches to address the challenges for *Salmonella* serotyping and microbial source tracking (MST). Firstly, a web-based bioinformatics tool to predict *Salmonella* serotypes (SeqSero, http://www.denglab.info/SeqSero) was developed by identifying the *Salmonella* antigen determinants (e.g. *rfb* gene cluster, *wzx*, *wzy*, *fliC* and *fljB* genes) from WGS data. Based on our curated databases for the determinants, SeqSero can theoretically achieve almost full spectrum *Salmonella* serotyping. Three datasets were used to evaluate the performance of SeqSero: 1) 308 raw reads genomes from *Salmonella* isolates of known serotype which were confirmed by CDC; 2) 3,306 raw reads genomes from *Salmonella* isolates sequenced and made publicly available by GenomeTrakr, a U.S. national monitoring network operated by the Food and Drug Administration; and 3) 354 other publicly available draft or complete *Salmonella* genomes from NCBI. The evaluation showed that SeqSero can reliably predict serotypes from WGS data with a high accuracy and fast speed. Secondly, we analyzed the population structure of a broad-host-range pathogen *Salmonella enterica* serovar Typhimurium (ST). A total of 1,267 ST genomes from clinical and various animal, food and environmental sources were included to identify

population groups (major phylogenetic lineages) and clades (smaller phylogenetic groups within a group) as well as their association with particular sources. A maximum likelihood phylogenetic tree was constructed based on whole genome sequencing SNPs (wgSNPs). A total of 10 major population groups were identified. Clustering of isolates from the same source was observed in 6 population groups, including clusters overrepresented by isolates from poultry, bovine, swine, and wild birds. Analyses of evolutionary relationship, metabolic profile, gene contents and pseudogene distribution provided further support for the source-cluster association. The observed source association demonstrated the potential of WGS-based subtyping in MST for ST, which was initially evaluated and analyzed by two different sets of genomes, one from publicly available genomes in the FDA GenomeTrakr database.

INDEX WORDS: serotyping, SeqSero, whole genome sequencing, microbial source tracking, pseudogene, *Salmonella enterica* serovar Typhimurium, population structure

SALMONELLA SEROTYPING AND SOURCE PREDICTION USING WHOLE GENOME SEQUENCING DATA

by

SHAOKANG ZHANG

B.S., Jilin University, China, 2010

M.S., China Agricultural University, China, 2013

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2017

SALMONELLA SEROTYPING AND SOURCE PREDICTION USING WHOLE GENOME SEQUENCING DATA

by

SHAOKANG ZHANG

Major Professor:	Dr. Xiangyu Deng
Committee:	Dr. Mark A. Harrison
	Dr. Ynes Ortega
	Dr. Patricia Fields

Electronic Version Approved:

Suzanne Barbour
Dean of the Graduate School
The University of Georgia
May 2017

DEDICATION

To Shangci Wang

my beautiful and amazing wife,

whose unselfish love and support made it possible for me to complete this work

and to my parents,

Wenxiang Zhang, and Ling Wang,

who make me be who I am

ACKNOWLEDGEMENTS

Firstly, I would like to thank my major advisor, Dr. Xiangyu Deng, who guided me to the area of bioinformatics and to finish this thesis. To me, he is half a mentor and half a good friend, who is always willing to help me when I meet difficulties and confusions. I also appreciate my committee members, Dr. Patricia Fields, Dr. Mark A. Harrison, and Dr. Ynes Ortega for their guidance and suggestions on my research projects. Thanks to my labmates Yan Qi and Shaoting Li for their help in both my research and my daily life. I am also grateful to Dr. Lee Katz, Blake Dinsmore and Charlotte Steininger from Enteric Diseases Laboratory Branch in Centers for Disease Control for their support to my research work. Special thanks to Wayne Harvester, Yanlong Yin and Zhenzhen Zhang, who helped me by providing reliable IT or statistics support. At last, I would also like to thank all my friends and the staff in both the Center for Food Safety and Department of Food Science and Technology of University of Georgia. I couldn't make it without all your support. Really appreciate it.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION

Salmonella, as one of the most frequently reported foodborne pathogens, is a major food safety and public health concern globally, causing as many as 1.3 billion clinical cases of disease annually worldwide (1). This pathogen can spread by various food vehicles such as poultry, eggs, beef, and milk products, etc. (2). To control its risk, multiple approaches have been applied in the field of food safety and public health.

Among them, serotyping forms the basis of current *Salmonella* surveillance systems in United States. The traditional serotyping is based on the immunologic agglutination between *Salmonella* cell antigens and specific antisera, which is time-consuming, labor-intensive and logistically challenging in preparation of the large set of full-spectrum antisera. Thus molecular based methods were developed based on both indirect targets, i.e. genetic markers or subtyping patterns associated with particular serotypes (3-6), and direct targets, i.e. the genetic determinants of serotypes i.e. *rfb* gene cluster (7, 8), *fliC (9)* and *fljB* (10) genes were both developed. The latter one has the advantage that it maintains the direct continuity between the phenotypic and genotypic determination of serotypes (11, 12) without the need to validate association relationships between markers and serotypes. With the trend that WGS is poised to transform approaches in clinical and public health microbiology (13), multiple public health microbiology related approaches integrated to single WGS platform, such as to predict multi-locus sequence typing (14, 15), to determine antimicrobial resistance genes (16) and to identify virulence

profiles (17). Thus to link WGS with the first line subtyping method in *Salmonella*, i.e. serotyping can be necessary.

Microbial source tracking (MST) methods can help to identify pathogen contamination sources which has been mainly applied in water system related investigations. Multiple studies have utilized MST or quasi-MST methods to estimate the burden of food safety and trace back the origins of foodborne pathogens (18, 19). However, numerous challenges, such as low discriminatory power, lack of standardized and promulgated method still remain in current MST when facing a set of different environments and situations (20). The application of WGS to MST has already been proposed to address some of the challenges (20). But so far, the study directly applied WGS-based subtyping in MST is yet to be performed. The major pre-condition for WGS-based MST is that the association relationships should be observed between specific sources and specific phylogeny population groups. Thus, it is necessary to investigate the population structure first, before applying WGS-based MST for one specific pathogen.

This dissertation is divided into 5 chapters. The first chapter presents an introduction and rationale on which the dissertation is based. The second chapter provides the literature review on topics including *Salmonella* serotyping, WGS, WGS-based subtyping and MST. The third chapter described SeqSero, a web-based bioinformatics tool, SeqSero (http://www.denglab.info/SeqSero) aiming to predict serotypes from WGS data. The fourth chapter investigated the population structure of sampled ST isolates, and demonstrated the association between certain ST population groups and specific sources, based on which an initial evaluation of MST for ST was performed. The final chapter is the summary of this dissertation which outlined our overall conclusions.

References

1. Bhunia A. 2007. Foodborne microbial pathogens: mechanisms and pathogenesis. Springer Science & Business Media.
2. Foodsafety.gov. (n.d.) *Salmonella* [Web log post]. Retrieved January 04, 2017, from https://www.foodsafety.gov/poisoning/causes/bacteriaviruses/*Salmonella*/index.html.
3. Zou W, Lin WJ, Foley SL, Chen CH, Nayak R, Chen JJ. 2010. Evaluation of pulsed-field gel electrophoresis profiles for identification of *Salmonella* serotypes. J Clin Microbiol 48:3122-3126.
4. Zou W, Lin WJ, Hise KB, Chen HC, Keys C, Chen JJ. 2012. Prediction system for rapid identification of *Salmonella* serotypes based on pulsed-field gel electrophoresis fingerprints. J Clin Microbiol 50:1524-32.
5. Kotetishvili M, Stine OC, Kreger A, Morris JG, Jr., Sulakvelidze A. 2002. Multilocus sequence typing for characterization of clinical and environmental *Salmonella* strains. J Clin Microbiol 40:1626-35.
6. Wise MG, Siragusa GR, Plumblee J, Healy M, Cray PJ, Seal BS. 2009. Predicting *Salmonella enterica* serotypes by repetitive sequence-based PCR. J Microbiol Methods 76:18-24.
7. Samuel G, Reeves P. 2003. Biosynthesis of O-antigens: genes and pathways involved in nucleotide sugar precursor synthesis and O-antigen assembly. Carbohydr Res 338:2503-19.
8. Jiang XM, Neal B, Santiago F, Lee SJ, Romana LK, Reeves PR. 1991. Structure and sequence of the *rfb* (O antigen) gene cluster of *Salmonella* serovar Typhimurium (strain LT2). Mol Microbiol 5:695-713.
9. Smith NH, Selander RK. 1990. Sequence invariance of the antigen-coding central region of the phase 1 flagellar filament gene (*fliC*) among strains of *Salmonella* typhimurium. J Bacteriol 172:603-9.
10. Vanegas RA, Joys TM. 1995. Molecular analyses of the phase-2 antigen complex 1,2,.. of *Salmonella* spp. J Bacteriol 177:3863-4.

11. Fitzgerald C, Collins M, van Duyne S, Mikoleit M, Brown T, Fields P. 2007. Multiplex, bead-based suspension array for molecular determination of common *Salmonella* serogroups. J Clin Microbiol 45:3323-3334.

12. McQuiston JR, Waters RJ, Dinsmore BA, Mikoleit ML, Fields PI. 2011. Molecular determination of H antigens of *Salmonella* by use of a microsphere-based liquid array. J Clin Microbiol 49:565-573.

13. Robilotti E, Kamboj M. 2015. Integration of whole-genome sequencing into infection control practices: the potential and the hurdles. J Clin Microbiol 53:1054-5.

14. Larsen MV, Cosentino S, Rasmussen S, Friis C, Hasman H, Marvig RL, Jelsbak L, Sicheritz-Ponten T, Ussery DW, Aarestrup FM, Lund O. 2012. Multilocus sequence typing of total-genome-sequenced bacteria. J Clin Microbiol 50:1355-61.

15. Inouye M, Conway TC, Zobel J, Holt KE. 2012. Short read sequence typing (SRST): multi-locus sequence types from short reads. BMC Genomics 13:338.

16. Zankari E, Hasman H, Cosentino S, Vestergaard M, Rasmussen S, Lund O, Aarestrup FM, Larsen MV. 2012. Identification of acquired antimicrobial resistance genes. J Antimicrob Chemother 67:2640-4.

17. Joensen KG, Scheutz F, Lund O, Hasman H, Kaas RS, Nielsen EM, Aarestrup FM. 2014. Real-time whole-genome sequencing for routine typing, surveillance, and outbreak detection of verotoxigenic *Escherichia coli*. J Clin Microbiol 52:1501-10.

18. Nayak R, Stewart-King T. 2008. Molecular epidemiological analysis and microbial source tracking of *Salmonella enterica* serovars in a preharvest turkey production environment. Foodborne Pathog Dis 5:115-26.

19. Cooley M, Carychao D, Crawford-Miksza L, Jay MT, Myers C, Rose C, Keys C, Farrar J, Mandrell RE. 2007. Incidence and tracking of *Escherichia coli* O157:H7 in a major produce production region in California. PLoS One 2:e1159.

20. Sadowsky MJ, Call DR, Santo Domingo JW. 2007. The future of microbial source tracking studies, p 235-277. *In* Santo Domingo J, Sadowsky M, Doyle M (ed), Microbial Source Tracking. American Society of Microbiology, Washington, DC.

CHAPTER 2

LITERATURE REVIEW

1. *Salmonella* in food safety and public health

Salmonella, as one of the most frequently reported foodborne pathogens, is a major food safety and public health concern globally, causing as many as 1.3 billion clinical cases of disease annually worldwide (1). In United States, it is estimated that about one million cases of foodborne illnesses are caused by *Salmonella* (with about 19,000 hospitalizations and 380 deaths) each year (2). *Salmonella* can be divided into two species: *Salmonella enterica* and *Salmonella bongori*. And in *Salmonella enterica*, there are six subspecies: *Salmonella enterica* subsp. *enterica*, *Salmonella enterica* subsp. *salamae*, *Salmonella enterica* subsp. *arizonae*, *Salmonella enterica* subsp. *diarizonae*, *Salmonella enterica* subsp. *indica*, and *Salmonella enterica* subsp. *houtenae*. Altogether they include over 2,500 serotypes (3). *Salmonella enterica* subsp. *enterica* has the vast majority of recognized serotypes (more than 2,300) and is responsible for almost all warm blood animal infections. According to national enteric diseases surveillance on *Salmonella* in 2013 by Centers for Disease Control and Prevention (CDC), Enteritidis and Typhimurium were the most frequently isolated serotypes followed by Newport, Javiana and Heidelberg (4).

Salmonella typically invades human and animal intestinal tract and spreads out of the host by shedding through feces. Generally, enteric fever (typhoid), diarrhea (enterocolitis), bacteremia and chronic asymptomatic carriage are the four major clinical presentations of salmonellosis (5). Among them, diarrhea is most common. And most people showing diarrhea symptoms between

12 and 72 hours after infection. And most infected individuals recover without treatment after 4 to 7 days. However, if the *Salmonella* infection spreads from intestines to other body organs and causes extraintestinal illness such as enteric fever (i.e., typhoid and paratyphoid fever) and invasive nontyphoidal *Salmonella* infection, specific treatments (e.g. antibiotics) should be provided (6, 7). Specific populations such as children, the elderly and people with weakened immune systems have higher risks of contracting *Salmonella* (8). Especially for children, from CDC's recent records, they are at the highest risk (9).

2. *Salmonella* virulence

Salmonella virulence is mainly determined by so-called virulence factors (10). It's hard to strictly define the term virulence factor (or virulence gene) (11). In general, bacterial virulence factors can be regarded as factors that enable a pathogen to replicate and disseminate within a host in part by subverting or eluding host defenses (12). The majority of virulence factor genes of *Salmonella* are clustered in *Salmonella* pathogenicity islands (SPI), which are genomic islands integrated into the chromosome of *Salmonella* (13, 14). Previous research showed that SPIs had been acquired by *Salmonella* from other species or sources through horizontal gene transfer (15, 16). So far, more than 14 SPIs have been identified in the *Salmonella* genome (17). Besides SPI, another major source of virulence factor genes is virulence plasmid. For example, *Salmonella enterica* serotype Abortusovis, Choleraesuis, Dublin, Enteritidis, Gallinarum/Pullorum, and Typhimurium are known to harbor a virulence plasmid with *Salmonella* plasmid virulence (spv) locus, which plays an important role in the in the intracellular multiplication of *Salmonella*e (18).

Bacterial secretion systems are highly involved in the virulence expression of *Salmonella*. They are utilized to transport virulence factors across cell walls from bacterial cells to host cells (19). More than seven different types of bacterial secretion systems have been described based

on their structures, functions, and specificity (20). Among them, type three secretion system (T3SS) which forms syringe-like organelles to translocate virulence factors is the most important one to *Salmonella* pathogenesis. Those virulence factors translocated by T3SS are aimed to modify the structure or biology (actin, formin, etc.) of the host cells to ensure a successful infection. Two T3SSs are commonly used by *Salmonella*: T3SS-1 encoded by SPI1 and T3SS-2 encoded by SPI2 (Figure 3.1) (14). The former one is mainly used in the intestinal invasion and onset of diarrhea, while the latter one plays a role in intracellular survival, replication and systemic infection (21, 22). It had long been thought that the two T3SSs functioned separately without overlap. But growing evidence suggests that the two T3SSs seem to be regulated together by the same mechanism with their action time overlapped with each other (23).

3. *Salmonella* Surveillance and Subtyping

3.1 *Salmonella* surveillance

According to Noordhuizen, et al. (24), surveillance refers to "a specific extension of monitoring where obtained information is utilized, and measures are taken if certain threshold values related to disease status have been passed". Since *Salmonella* is widely distributed in the environment and can enter the food chain through multiple points by multiple routes (25), the establishment of an efficient and broad surveillance system is highly essential. Although not a control measure itself, surveillance is an important adjunct to any control systems designed to prevent or minimize salmonellosis (26). *Salmonella* surveillance can be used not only in direct disease prevention and control, such as to provide the warnings of hazards, the magnitude of outbreak, locations of outbreak, effectiveness of correction, explanation of outbreak sources, etc. (26), but also in some other fields such as risk analysis and risk management (27). Most *Salmonella* surveillance systems are based on *Salmonella* subtypes (which will be discussed in

details in later part). For example, *Salmonella* serotyping forms the basis of the US National *Salmonella* Surveillance System (28). PulseNet, a national network of laboratories for foodborne illness surveillance applies pulsed field gel electrophoresis (PFGE) and multiple locus variable number tandem repeat analysis (MLVA) subtyping techniques for outbreak detections (29, 30).

In the United States, multiple surveillance systems work together to collect *Salmonella* infection data for CDC (31, 32) with different targets and approaches:

The National *Salmonella* Surveillance System is a surveillance system collecting data through passive surveillance for laboratory-confirmed human *Salmonella* isolates. *Salmonella* isolates are isolated from clinical laboratories (such as hospitals) and then submitted to state and territorial public health laboratories for confirmation and serotyping. If they are found to be uncommon serotypes or un-typeable isolates, the samples will be sent to National *Salmonella* Reference Laboratory (NSRL) at the Enteric Diseases Laboratory Branch (EDLB) in CDC for further characterization or confirmation; the final reports about those strains will be reported back to state and territorial public health labs. The summaries of data are released annually (33).

A foodborne disease outbreak occurs when two or more people get the same illness from the same contaminated food or drink (34). Foodborne Disease Outbreak Reporting System (FDOSS) is designed to collect foodborne outbreak information from state, local and territorial public health agencies who identify and investigate such outbreaks. A web-based, electronic system (eFORS) has been created based on FDOSS (https://wwwn.cdc.gov/foodborneoutbreaks/).

The National Antimicrobial Resistance Monitoring System for Enteric Bacteria (NARMS) is a surveillance system established in 1996 to monitor and track antimicrobial resistance of pathogens such as *Salmonella*, *Escherichia coli* (*E. coli*), *Campylobacter* and *Shigella*. NARMS

is an interagency partnership among state and local public health departments, CDC, the U.S. Food and Drug Administration (FDA), and the U.S. Department of Agriculture (USDA).

The Foodborne Diseases Active Surveillance Network (FoodNet) is another important surveillance system. The major targets of FoodNet are to determine, monitor and sources-attribute the burden of foodborne illness over time, which provides a foundation for developing national food safety guidelines, goals and control measures. Similar to NARMS, FoodNet is a corporation network by CDC, FDA, USDA and state health departments. It covers 10 states and 15% national population (35). As an active surveillance system, instead of local clinical labs reporting cases to FoodNet, FoodNet public health officials routinely communicate with surveillance area clinical laboratories for new cases. FoodNet is the only surveillance system that focuses on pathogen infections through food.

PulseNet is a national laboratory network that applies standardized molecular subtyping methods (such as PFGE) to detect local and multistate outbreaks (29). A national computer network is used to detect putative outbreaks of foodborne illness by linking cases occurring in same or different geological areas. Specifically, PulseNet participating labs across the nation submit DNA fingerprints of bacteria based on standard protocols from sick cases to CDC PulseNet databases. Fingerprints are then compared by looking for cases of matching patterns that surge beyond the baseline level of sporadic cases within a specific period of time. Such PulseNet analysis is usually the first step in identifying a widespread foodborne outbreak. Then epidemiological investigation will be launched to try to detect the source of the illnesses (30). Besides PulseNet at the national level, to face the challenge of international foodborne pathogen outbreaks favored by the increasing international trade and travel, PulseNet International has also been developed (36).

3.2 *Salmonella* subtyping

A bacterial subtype is to describe a set of isolates with common phenotypic or molecular characteristics based on which it can be distinguished from other isolates of the same species (37). *Salmonella* subtyping techniques have been playing a significant role in laboratory-based surveillance systems and outbreak investigations of *Salmonella* (38, 39). Based on the approaches and characteristics, pathogen subtyping can be classified as DNA based subtyping (molecular subtyping) and phenotype based subtyping (37).

3.2.1 *Salmonella* molecular subtyping

Molecular subtyping methods have been widely used for tracing bacterial pathogens associated with animal and human disease outbreaks (29, 30, 40). The general procedure of molecular subtyping methods is to compare the DNA genetic markers ("fingerprint") of two or more pathogen isolates to determine whether they belong to same "cluster" (37, 41). Some common molecular subtyping methods for *Salmonella* are:

Pulsed Field Gel Electrophoresis (PFGE). PFGE was firstly developed as a method for circumventing the limitations of conventional gel electrophoresis techniques which can't work with very large genomic restriction fragments (42, 43). The DNA, immobilized in the agarose, is cut with restriction enzymes at specific locations (restriction sites). Then an electric field that periodically changes direction will be applied to gel matrix for better separation of DNA fragments. PulseNet scientists can find outbreaks by comparing and connecting similar PFGE fingerprints (using a software program known as BioNumerics) of foodborne illness and linking these illnesses across states and countries. Now PFGE has been widely utilized in *Salmonella* and other foodborne pathogen outbreak investigation and considered as the "gold standard" for molecular subtyping (44, 45) due to the advantages of 1) high concordance with epidemiological

relatedness, 2) universal application to different pathogens, 3) good stability and reproducibility, etc.

Multiple Locus Variable number tandem repeat Analysis (MLVA). Tandem repeated sequences naturally occur in many different loci of bacteria chromosomal regions. This kind of variant is inheritable, which can be applied for personal or parental identification (46). Public health scientists can measure them by PCR and take advantage of the polymorphism of the tandemly repeated DNA sequences to differentiate suspected and fast-evolving bacterial strains from an outbreak (47). In PulseNet surveillance system, MLVA can help to further differentiate subtypes during outbreak investigations after PFGE. PFGE and MLVA data should be used in conjunction with epidemiological evidence and other outbreak related information from investigations to detect an outbreak.

Multi-Locus Sequence Typing (MLST). MLST is a molecular subtyping approach which measures the variability in small parts of the genome (5 to 7 house-keeping genes based on individual MLST scheme) caused by the house-keeping genes' mutation or recombination events. Based on the designation of MLST, the number of nucleotide differences between alleles is ignored and sequences are given different allele numbers whether they differ at a single nucleotide site or at many sites (48, 49). A well-curated scheme is needed for MLST analysis. It has been used to investigate population structures and discriminate epidemiological clones of bacterial pathogens (50). However, compared to other molecular subtyping techniques, traditional MLST is less discriminating, more expensive and time-consuming (51).

In the recent decades, the application of molecular subtyping has greatly improved our ability to distinguish epidemiologically related isolates. However, even with the assistance of molecular subtyping methods, the discriminatory power is still not enough for some highly clonal pathogen

populations. For example, among the *Salmonella enterica* Serotype Enteritidis isolates reported to PulseNet (29), JEGX01.0004 accounts for approximately 45% of strains, making it difficult to further differentiate isolates that share this common PFGE type to identify outbreaks. New methods with higher discriminatory power are needed for such situations.

3.2.2 *Salmonella* serotyping

Compared with DNA based subtyping, phenotypic subtyping or phenotyping is based on phenotypic characteristics. Some common phenotyping methods for *Salmonella* are: serotyping which is based on immunologic reactivity between antisera and *Salmonella* surface antigens (the agglutination of antibody and antigen), phage typing which is based on the susceptibility between *Salmonella* strains and a standard set of phages, and multi-locus enzyme electrophoresis which is based on variations of different strains in electrophoretic mobility of different constitutive enzymes (37). The limited discriminatory power limited the application of phenotyping (37). However, certain phenotyping methods are still commonly used. For example, serotyping continues to be the first line subtyping method for *Salmonella* surveillance, (28). It has become an important approach for detecting *Salmonella* outbreaks, monitoring trends of disease, and attributing sources of salmonellosis since 1960s (52, 53). For more than 50 year, serotyping serves as the foundation of the US National *Salmonella* Surveillance System.

The key determinants of *Salmonella* serotypes are lipopolysaccharide O antigen and flagellar H antigens (two H antigens: H1 antigen and H2 antigen). There are 46 recognized O antigens and 114 different H antigens described in the Kauffmann–White–Le Minor scheme for *Salmonella* serotype identification. By different combinations of O and H antigen types, more than 2,500 different serotypes have been identified and described in the White-Kauffmann-Le Minor scheme (3). The O antigen is encoded by the *rfb* gene cluster (54). The flagellar H1 and

H2 antigens are encoded by *fliC* and *fljB* genes respectively (55). There is a set of genes in the *rfb* gene cluster involved in O antigen biosynthesis. The genetic variation of the *rfb* gene cluster is related to O antigen's functions in mediating selective recognition by host immune system and bacteriophages (56). In addition, O antigen was proposed to be involved in *Salmonella* virulence (57). For example, O-antigen mutations of *Salmonella enterica* serovar Typhimurium can cause attenuated ability to colonize chick/calf intestines and attenuated clinical phenotypes (58). The O-antigen also provided protection to *Salmonella enterica* serovar Typhi and enhanced its survival in human serum (59). The *fliC* and *fljB* alleles share sequence similarity in their 5' and 3' ends and display variations in the middle part of the genes (55). The *fliC* gene is located in the flagellar biosynthesis operon present in all enteric bacteria, whereas the *fljB* gene is unique to *Salmonella* (60, 61). The two genes are coordinately regulated by a phase variation mechanism and only one gene can be expressed in a single cell at one time (62).

The traditional serotyping method is labor intensive and time-consuming (taking at least 2 days). It requires production and quality control of multiple antisera, which can be logistically challenging if the capacity for identifying many serotypes is needed. Thus molecular *Salmonella* serotyping methods have been developed based on two strategies. First, genetic markers perceived to be associated with particular serotypes, such as PFGE patterns (63, 64), MLST profiles (65) and repetitive sequence-based PCR patterns (66), have been proposed to predict serotypes. Even though such methods have shown various degrees of success in serotype prediction, their accuracy can be compromised by horizontal transfer of either the surrogate markers or the genetic determinants of serotype. Second, the genetic determinants of *Salmonella* serotypes i.e. *rfb* gene cluster, *fliC* and *fljB* genes which have been directly targeted for serotype determination. For example, Fitzgerald et al. developed a Bio-Plex platform based assay for the

detection of six most common O antigen groups in the United States (B, C1, C2, D, E, and O13) plus serotype Paratyphi A (67), and McQuiston et al. extended the assay to the determination of the common H antigens (68). This method maintains the important continuity between the phenotypic and genotypic determination of serotypes (60, 69).

4. Whole genome sequencing (WGS) and WGS-based subtyping

4.1 WGS Basics

WGS is a laboratory process to determine the entirety DNA information of an organism's genome. Traditional sequencing methods (e.g., Sanger sequencing) are time consuming, low throughput and expensive. The emergence of high-throughput sequencing technologies (also called Next Generation Sequencing, NGS), which allow to sequence millions of DNA sequences concurrently in a single next-generation sequencing process, has transformed WGS into a more convenient, lower-cost and faster procedure (70, 71). For example, it takes less than 48 hours to sequence 61 *Salmonella* genomes at an average of 50 X sequencing coverage in a single Illumina MiSeq run based on Illumina sequencing coverage calculator (calculated on January 15, 2017).

Illumina, Ion Torrent and Pacific Biosciences (PacBio) are three current major platforms for WGS. Illumina and Ion Torrent are short-reads sequencing technologies which generate millions of short reads (up to several hundred nucleotides per read). Illumina applies a so called sequencing by synthesis (SBS) approach (72), with DNA sequence determined by the type and order of the four fluorescently-labeled nucleotides (A, T, G, C) that are incorporated into the complementary strand of a genome fragment being sequenced. The SBS process occurs in millions of clusters of DNA templates prepared from one or multiple genomes in parallel, allowing high throughout readouts of short sequences, i.e., sequencing reads, to be generated.

The Illumina platform has become the most widely used WGS technology, accounting for over 90% of DNA sequences deposited in public databases (73).

Ion Torrent sequencing technology is also called Ion semiconductor sequencing. The small DNA fragments with specific adapters are linked to the surface of Ion Sphere Particles (3-micron diameter beads) then amplified by emulsion PCR. Subsequently, they will be loaded to proton sensing wells where the sequencing proceeds. Four bases will be introduced sequentially, if the nucleotide base matches, a proton will be released to change pH. This pH change will then be sensed by a sensitive pH sensor to output the sequences of the reads. Compared to Illumina, Ion Torrent has lower instrument cost and faster run, but with a higher error rate (74).

In comparison with Illumina and Ion Torrent, PacBio features the output of much longer sequencing reads, averaging over 10,000 nucleotides in length. PacBio sequencing relies on a process that enables single molecule real time (SMRT) sequencing (75). It utilizes 50 nm wide wells which are called zero-mode waveguides (ZMWs). These wells are small enough for an illuminated observation of only a single nucleotide incorporation. DNA polymerase is affixed in the bottom of ZMWs with single DNA molecule and four phospho-labeled dNTP added. As each base is incorporated by polymerase, the fluorescent tag is cleaved off to create a distinctive pulse of fluorescence which can be caught by detectors in real time. Then the fluorescence will disappear when the tag diffuses out of observation area. PacBio produces long reads with relatively low reads accuracy (higher error rate); however, post-run bioinformatics analysis like consensus alignment can help to increase its accuracy (76).

Besides the aforementioned three sequencing techniques, nanopore sequencing has been rapidly evolving to become a promising new platform for WGS (77). A nanopore is simply a small hole with an internal diameter of the order of one nanometer. When an electrical potential

is applied across the membrane of nanopores, electronic current flows occur through the aperture. When nucleotides pass through the nanopores, each different base disrupts the current to a different extent. This process allows each nucleotide to be identified by detecting changes of current. Nanopore sequencing has the advantages of long sequencing reads and a small footprint (hand-held) device. But compared to other platforms, its error rates are still relatively higher than other short reads platforms, which is to some extent similar to Pacbio (78). Efforts have been made to increase the accuracy and application of nanopore sequencing (79, 80).

4.2 WGS-based subtyping

WGS allows genome wide identification of sequence polymorphisms for subtyping. When phylogenetically informative markers such as single nucleotide polymorphisms (SNPs) are used, subtyping is built upon the reconstruction of phylogeny of analyzed isolates. This allows high-resolution differentiation of closely-related isolates by resolving their evolutionary relationship.

WGS-based subtyping has shown great potential to be applied in the area of food safety and public health (70, 71, 81). Compared with other molecular subtyping methods, the biggest advantage of WGS-based subtyping is the highly improved discriminatory power (82). The first application of WGS in a foodborne outbreak investigation was reported in 2010 (83). During a large-scale listeriosis outbreaks caused by ready-to-eat meat products (Canada, 2008), WGS was performed to analyze of two outbreak-associated isolates that displayed similar but distinct PFGE patterns. It led to the conclusion that multiple distinct but highly related strains may have been involved in the outbreak, which could not have been be achieved by traditional subtyping methods.

Since then, WGS-based methods have been increasingly applied in foodborne pathogen surveillance and outbreak investigation. For example, during the investigation of a multistate

Salmonella enterica serotype Montevideo outbreak that implicated contaminated black and red peppers for production of Italian-style deli meats (84), Lienau et al. used WGS-based subtyping when PFGE failed to distinguish isolates of this outbreak from those of previous *Salmonella. enterica* serotype Montevideo outbreak caused by contaminated pistachios (85). WGS provided convincing evidence to link clinical samples to a food manufacture facility. This is regarded as the first significant application of WGS subtyping in outbreak source identification that led to regulatory decision making (72).

Another example is the application of WGS in the investigation of the 2011 Shiga toxin – producing *Escherichia coli* O104:H4 outbreak in Germany (86, 87). Caused by contaminated sprouts, this was one of the largest foodborne outbreak in recent history that led to around 4,000 cases of bloody diarrhea, 850 cases of hemolytic uremic syndrome and 50 deaths (86). The investigation of this outbreak featured the real-time release of WGS data and crowdsourced analyses by investigators from multiple countries. In less than a week, the collaborative study revealed that the outbreak strain had evolved from enteroaggregative *E. coli* (EAEC) by the acquisition of virulence factors specific to enteroaggregative hemorrhagic *E. coli* (EHEC), such as *Stx2* genes (87).

There are two major methods for WGS-based subtyping: whole genome single nucleotide polymorphisms (wgSNP) typing and whole genome multi-locus sequence typing (wgMLST).

Whole genome SNPs are a set of genome-wide single nucleotide mutations that can be used to infer phylogenetic relationship among closely-related isolates. The first application of wgSNP in high-throughput subtyping of a bacterial pathogen was reported in 2010 by Harris, et al. (88). In this seminal study on methicillin-resistant *Staphylococcus aureus* (MRSA), raw sequencing reads from the Illumina platform were aligned to a reference genome to identify high quality

wgSNPs between the reference isolate and a query isolate. A phylogenetic tree was then created using the wgSNPs. This approach revealed the geographic structure of this pathogen on a global scale and provided sufficient resolution to identify transmission both between continents and within a local hospital. This study laid the foundation for the WGS-based subtyping using wgSNP. Since then, wgSNPs have been widely used in both phylogenetic studies and outbreak investigations. For example, population structure of prevalent populations of *Salmonella enterica* Serotype Enteritidis in the United States was investigated by wgSNP typing (82). In the investigation of a *Listeria* outbreak, wgSNP typing results allowed differentiation between outbreak cases and sporadic cases, which assisted further epidemiological investigation (89).

While wgSNP typing has the advantages of high-resolution in differentiating closely-related isolates and allowing robust evolutionary relationship to be inferred (72), some shortcomings have been identified for its application in food safety and public health surveillance. For example, since the exact set of wgSNPs is subjected to change when different query or reference genomes are selected for analysis, a standard subtype nomenclature is difficult to form for subtyping using wgSNP Also, wgSNP subtyping requires an appropriate reference genome that is preferably fully assembled and closely-related to genomes to be subtyped. However, such an optimal reference is not always available.

To avoid these disadvantages, wgMLST, a WGS-based upgrade of tradition MLST scheme has been recently proposed (72, 90). While MLST is very useful for studying population structures, it cannot often achieve strain level differentiation using sequences from only a few housekeeping genes. Therefore, it only has limited application in outbreak investigation (91). With WGS, it is possible to expand traditional MLST to wgMLST through interrogating thousands of loci across entire genomes (92). Since many more loci will be included in a

wgMLST scheme, a much higher discriminatory power can be achieved. The first application wgMLST was reported in an investigation of the ecology and population structure of *Campylobacter.* A scheme set of 1,643 well annotated and defined loci was used to afford sufficient discriminatory power to identify outbreaks and trace to (90) potential sources of the pathogen. Since then, wgMLST has been incorporated into actual outbreak investigations and surveillance. For example, Kovanen, et al. applied wgMLST in the surveillance of *Campylobacter jejuni* infections during a seasonal peak in Finland (93). Ruppitsch et al. developed a *Listeria monocytogenes* scheme containing 1,701 loci from 42 *L. monocytogenes* genomes (36 complete RefSeq genomes and 6 draft genomes) (94). Chen et al. not only showed that wgMLST can be applied to differentiate outbreak strains and unrelated strains for *L. monocytogenes* investigation, but also extended the study by developing a specific scheme for each major lineage of *L. monocytogenes,* which provided improved discriminatory power and epidemiological concordance (95). Despite some known issues such as including the requirement of highly curated schemes, underestimation of actual genetic variations when multiple mutations in a gene are represented as a single allele , and lower resolution than wgSNP subtyping, wgMLST is regarded as a promising mainstream method for public health surveillance and outbreak investigation of foodborne pathogens (72).

5. Microbial source tracking

A variety of sources of hazards can lead to foodborne diseases. To improve food safety and public health, it is important to identify and determine the sources of foodborne contamination. Microbial source tracking (MST), which has been mainly used to help identify sources responsible for the fecal pollution of water systems, is the process of tracing fecal pathogens to their origin using genotypic and phenotypic methods (96). It has great potential to be applied to

food safety and public health area due to its ability to provide information on where to focus, allocate resources and intervene in the food production chain (96, 97).

In 1960s, early MST attempts included the use of fecal coliform-to-fecal streptococci ratios (US EPA guide 2005), which is now regarded as an uninformative method. Since then, a number of innovative MST techniques have been proposed and evaluated (98). Even though with decades of development, MST is still an emerging field facing multiple challenges.

Currently, MST methods can be commonly classified into two categories: library dependent methods (LDMs) and library independent methods (LIMs). LDMs require a large library of phenotypic and genotypic fingerprints (or characteristics) of known pathogen sources to compare with the tested isolates. For example, resistance profiles, carbon utilization profiles, molecular subtypes (PFGE, rep-PCR, MLST, etc.), etc. (96, 97, 99) Due to the effort it takes to develop a library, LDMs tend to be more expensive and time-consuming (100). Compared to LDMs, LIMs don't depend on fingerprint libraries but rely on particular genotypic traits to identify sources of pathogens. The most common LIM approach is host-specific molecular marker PCR, which is based on the detection of a specific host associated genetic marker specifically associated with a host population (97). LIMs are more efficient and cost-effective, but the discriminatory power is relatively low (97). According to comparison studies, no single method is clearly superior to the others (US EPA guide 2005). The application of multiple methods might be a good strategy due to the lack of a standardized approach (96, 97, 99). In the future, additional development is needed for innovative MST techniques with features of increased sensitivity and specificity, real-time analysis, suitable discriminatory power, etc. (96, 97, 99, 101)

Multiple studies have utilized MST or quasi-MST methods to estimate the burden of food safety and trace back the origins of foodborne pathogens (102, 103). However, multiple

challenges still exist for MST, such as the requirement of large libraries, low rates of correct classification, low discriminatory power, lack of standardized and promulgated method, etc. (97, 98). Also commonly used molecular subtyping methods for bacterial pathogens such as PFGE and MLVA are of limited use in identifying pathogens sources such as their animal hosts (104). The reason is partly due to the lack of correlation between phenotypic or genotypic subtypes of the pathogens and their hosts.

The application of WGS to MST has already been proposed to address some of the challenges (104). But so far, to our knowledge, there has been only one published work that applied WGS to track the source of *E. coli* (105). In this study only a limited number (52 isolates) of *E. coli* isolates were sequenced in an attempt to identify source-specific markers for tracking *E. coli* contamination. To evaluate the performance of WGS in MST, a large sample size with genomes from various sources is essential.

6. Summary

Salmonella is one of the most common causes of food poisoning in the United States. To control its risk, multiple approaches have been applied in the field of food safety and public health. Among them, serotyping forms the basis of current *Salmonella* surveillance system. However, there are limitations in current serotyping approaches. For example, traditional *Salmonella* serotyping method based on antigen-antiserum reaction time-consuming and logistically challenging. MST methods can help to identify contamination sources which has been majorly applied in water system contamination related investigations. But numerous challenges, such as low discriminatory power, lack of standardized and promulgated method, still remain (104). Compared with conventional methods, WGS based analysis promised great potential for solving the challenges (106, 107).

Reference

1. Bhunia A. 2007. Foodborne microbial pathogens: mechanisms and pathogenesis. Springer Science & Business Media.
2. Scallan E, Hoekstra RM, Angulo FJ, Tauxe RV, Widdowson MA, Roy SL, Jones JL, Griffin PM. 2011. Foodborne illness acquired in the United States--major pathogens. Emerg Infect Dis 17:7-15.
3. Grimont PA, Weill FX. 2007. Antigenic formulae of the *Salmonella* serovars. WHO collaborating centre for reference and research on *Salmonella*.
4. CDC.gov. (n.d.) National Enteric Disease Surveillance: *Salmonella* Annual Report, 2013. Retrieved January 20, 2017, from https://www.cdc.gov/nationalsurveillance/pdfs/*Salmonella*-annual-report-2013-508c.pdf
5. Coburn B, Grassl GA, Finlay BB. 2007. *Salmonella*, the host and disease: a brief review. Immunol Cell Biol 85:112-8.
6. Crum-Cianflone NF. 2008. Salmonellosis and the gastrointestinal tract: more than just peanut butter. Curr Gastroenterol Rep. 10:424-431.
7. Andino A, Hanning I. 2015. *Salmonella enterica*: survival, colonization, and virulence differences among serovars. Scientific World J. 2015:520179.
8. Lund BM, O'Brien SJ. 2011. The occurrence and prevention of foodborne disease in vulnerable people. Foodborne Pathog Dis 8:961-973.
9. CDC.gov. (n.d.) Foodborne Diseases Active Surveillance Network (FoodNet): FoodNet Surveillance Report for 2014 (Final Report). Retrieved January 20, 2017, from https://www.cdc.gov/foodnet/pdfs/2014-foodnet-surveillance-report.pdf

10. van Asten AJ, van Dijk JE. 2005. Distribution of “classic” virulence factors among *Salmonella* spp. FEMS Immunol Med Microbiol 44:251-259.

11. Wassenaar TM, Gaastra W. 2001. Bacterial virulence: can we draw the line? FEMS Microbial Lett 201:1-7.

12. Cross AS. 2008. What is a virulence factor? Critical Care 12:196.

13. Schmidt H, Hensel M. 2004. Pathogenicity islands in bacterial pathogenesis. Clin Microbiol Rev 17:14-56.

14. Marcus SL, Brumell JH, Pfeifer CG, Finlay BB. 2000. *Salmonella* pathogenicity islands: big virulence in small packages. Microbes Infect 2:145-56.

15. Juhas M, van der Meer JR, Gaillard M, Harding RM, Hood DW, Crook DW. 2009. Genomic islands: tools of bacterial horizontal gene transfer and evolution. FEMS Microbiol Rev 33:376-93.

16. Hacker J, Kaper JB. 2000. Pathogenicity islands and the evolution of microbes. Annu Rev Microbiol 54:641-79.

17. Rhen M. 2007. *Salmonella*: molecular biology and pathogenesis. Horizon Scientific Press.

18. Rotger R, Casadesús J. 2010. The virulence plasmids of *Salmonella*. Int Microbiol 2:177-184.

19. Green ER, Mecsas J. 2016. Bacterial Secretion Systems–An overview. Microbiol Spectr. 4 (1).

20. Tseng T-T, Tyler BM, Setubal JC. 2009. Protein secretion systems in bacterial-host associations, and their description in the Gene Ontology. BMC Microbiol 9:S2.

21. Kaur J, Jain SK. 2012. Role of antigens and virulence factors of *Salmonella enterica* serovar Typhi in its pathogenesis. Microbiol Res 167:199-210.

22. de Souza Santos M, Orth K. 2015. Subversion of the cytoskeleton by intracellular bacteria: lessons from *Listeria*, *Salmonella* and Vibrio. Cell Microbiol 17:164-73.

23. Moest TP, Méresse S. 2013. *Salmonella* T3SSs: successful mission of the secret (ion) agents. Curr Opin Microbiol 16:38-44.

24. Noordhuizen JPTM, Frankena K, Thrusfield MV, Graat E. 2001. Application of quantitative methods in veterinary epidemiology. Wageningen Pers.

25. Ienistea C. 1967. The role of foods in the incidence of *Salmonella* infections. Microbiol Parazitol Epidemiol (Bucur) 12:127-33.

26. *Salmonella* NRCCo, Food US, Administration D. 1969. An evaluation of the *Salmonella* problem. National Academies.

27. Salman M. 2008. Animal disease surveillance and survey systems: methods and applications. John Wiley & Sons.

28. Robinson RK, Batt CA. 2014. Encyclopedia of food microbiology. Academic press.

29. Swaminathan B, Barrett TJ, Hunter SB, Tauxe RV, Force CDCPT. 2001. PulseNet: the molecular subtyping network for foodborne bacterial disease surveillance, United States. Emerg Infect Dis 7:382-9.

30. Boxrud D, Monson T, Stiles T, Besser J. 2010. The role, challenges, and support of pulsenet laboratories in detecting foodborne disease outbreaks. Public Health Rep 125:57-62.

31. Swaminathan B, Barrett TJ, Fields P. 2006. Surveillance for human *Salmonella* infections in the United States. J AOAC Int. 89:553-559.

32. Olsen SJ, MacKinnon LC, Goulding JS, Bean NH, Slutsker L. 2000. Surveillance for foodborne-disease outbreaks—United States, 1993–1997. MMwR CDC Surveill Summ 49:1-62.

33. CDC.gov. (2014, December 22). National *Salmonella* Surveillance. Retrieved January 30, 2017, from https://www.cdc.gov/nationalsurveillance/*Salmonella*-surveillance.html

34. Organization WH. 2008. Foodborne disease outbreaks: guidelines for investigation and control. World Health Organization.

35. Hardnett FP, Hoekstra RM, Kennedy M, Charles L, Angulo FJ. 2004. Epidemiologic issues in study design and data analysis related to FoodNet activities. Clin Infect Dis 38 Suppl 3:S121-6.

36. Swaminathan B, Gerner-Smidt P, Ng LK, Lukinmaa S, Kam KM, Rolando S, Gutiérrez EP, Binsztein N. 2006. 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 3:36-50.

37. Wiedmann M. 2002. Subtyping of bacterial foodborne pathogens. Nutr rev 60:201-208.

38. Hyeon JY, Chon JW, Park JH, Kim MS, Oh YH, Choi IS, Seo KH. 2013. A Comparison of Subtyping Methods for Differentiating *Salmonella enterica* Serovar Enteritidis Isolates Obtained from Food and Human Sources. Osong Public Health Res Perspect 4:27-33.

39. Deng X, Shariat N, Driebe EM, Roe CC, Tolar B, Trees E, Keim P, Zhang W, Dudley EG, Fields PI, Engelthaler DM. 2015. Comparative analysis of subtyping methods against a whole-genome-sequencing standard for *Salmonella enterica* serotype Enteritidis. J Clin Microbiol 53:212-8.

40. Adzitey F, Huda N, Ali GRR. 2013. Molecular techniques for detecting and typing of bacteria, advantages and application to foodborne pathogens isolated from ducks. Biotech 3:97-107.

41. Moorman M, Pruett P, Weidman M. 2010. Value and methods for molecular subtyping of bacteria, p 157-174, Principles of Microbiological Troubleshooting in the Industrial Food Processing Environment. Springer.

42. Schwartz DC, Cantor CR. 1984. Separation of yeast chromosome-sized DNAs by pulsed field gradient gel electrophoresis. Cell 37:67-75.

43. Schwartz D, Saffran W, Welsh J, Haas R, Goldenberg M, Cantor C. New techniques for purifying large DNAs and studying their properties and packaging, p 189-195. *In* (ed), Cold Spring Harbor Laboratory Press.

44. Foley SL, White DG, McDermott PF, Walker RD, Rhodes B, Fedorka-Cray PJ, Simjee S, Zhao S. 2006. Comparison of subtyping methods for differentiating *Salmonella enterica* serovar Typhimurium isolates obtained from food animal sources. J Clin Microbiol 44:3569-77.

45. Foley SL, Lynne AM, Nayak R. 2009. Molecular typing methodologies for microbial source tracking and epidemiological investigations of Gram-negative bacterial foodborne pathogens. Infect Genet Evol 9:430-40.

46. Khan FA. 2014. Biotechnology in medical sciences. CRC Press.

47. Weigel RM, Qiao B, Teferedegne B, Suh DK, Barber DA, Isaacson RE, White BA. 2004. Comparison of pulsed field gel electrophoresis and repetitive sequence polymerase chain reaction as genotyping methods for detection of genetic diversity and inferring transmission of *Salmonella*. Vet Microbiol 100:205-17.

48. Maiden MC, Bygraves JA, Feil E, Morelli G, Russell JE, Urwin R, Zhang Q, Zhou J, Zurth K, Caugant DA. 1998. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. Proc Natl Acad Sci U S A 95:3140-3145.

49. Urwin R, Maiden MC. 2003. Multi-locus sequence typing: a tool for global epidemiology. Trends Microbiol 11:479-487.

50. Harbottle H, White DG, McDermott PF, Walker RD, Zhao S. 2006. Comparison of multilocus sequence typing, pulsed-field gel electrophoresis, and antimicrobial susceptibility typing for characterization of *Salmonella enterica* serotype Newport isolates. J Clin Microbiol 44:2449-57.

51. Larsen MV, Cosentino S, Rasmussen S, Friis C, Hasman H, Marvig RL, Jelsbak L, Pontén TS, Ussery DW, Aarestrup FM. 2012. Multilocus sequence typing of total genome sequenced bacteria. J Clin Microbiol 50(4):1355-61.

52. Herikstad H, Motarjemi Y, Tauxe R. 2002. *Salmonella* surveillance: a global survey of public health serotyping. Epidemiol Infect 129:1-8.

53. Galanis E. 2006. Web-based Surveillance and Global *Salmonella* Distribution, 2000–2002. Emerg Infect Diseases 12(3):381-8.

54. Liu B, Knirel YA, Feng L, Perepelov AV, Sof'ya NS, Reeves PR, Wang L. 2014. Structural diversity in *Salmonella*O antigens and its genetic basis. FEMS microbiol rev 38:56-89.

55. McQuiston J, Parrenas R, Ortiz-Rivera M, Gheesling L, Brenner F, Fields P. 2004. Sequencing and comparative analysis of flagellin genes *fliC*, *fljB*, and *flpA* from *Salmonella*. J Clin Microbiol 42:1923-1932.

56. Reeves P, Wang L. 2002. Genomic organization of LPS-specific loci, p 109-135, Pathogenicity islands and the evolution of pathogenic microbes. Springer.

57. Liu B, Knirel YA, Feng L, Perepelov AV, Senchenkova SN, Reeves PR, Wang L. 2014. Structural diversity in *Salmonella* O antigens and its genetic basis. FEMS Microbiol Rev 38:56-89.

58. Morgan E, Campbell JD, Rowe SC, Bispham J, Stevens MP, Bowen AJ, Barrow PA, Maskell DJ, Wallis TS. 2004. Identification of host-specific colonization factors of *Salmonella enterica* serovar Typhimurium. Mol Microbiol 54:994-1010.

59. Kintz E, Heiss C, Black I, Donohue N, Brown N, Davies MR, Azadi P, Baker S, Kaye PM, van der Woude M. 2017. *Salmonella enterica* Serovar Typhi Lipopolysaccharide O-Antigen Modification Impact on Serum Resistance and Antibody Recognition. Infect Immun 85.

60. McQuiston JR, Waters RJ, Dinsmore BA, Mikoleit ML, Fields PI. 2011. Molecular determination of H antigens of *Salmonella* by use of a microsphere-based liquid array. J Clin Microbiol 49:565-573.

61. Macnab RM. 1992. Genetics and biogenesis of bacterial flagella. Annu Rev Genet 26:131-58.

62. Silverman M, Zieg J, Hilmen M, Simon M. 1979. Phase variation in *Salmonella*: genetic analysis of a recombinational switch. Proc Natl Acad Sci U S A 6:391-395.

63. Zou W, Lin W-J, Foley SL, Chen C-H, Nayak R, Chen JJ. 2010. Evaluation of pulsed-field gel electrophoresis profiles for identification of *Salmonella* serotypes. J Clin Microbiol 48:3122-3126.

64. Zou W, Lin WJ, Hise KB, Chen HC, Keys C, Chen JJ. 2012. Prediction system for rapid identification of *Salmonella* serotypes based on pulsed-field gel electrophoresis fingerprints. J Clin Microbiol 50:1524-32.

65. Kotetishvili M, Stine OC, Kreger A, Morris JG, Jr., Sulakvelidze A. 2002. Multilocus sequence typing for characterization of clinical and environmental *Salmonella* strains. J Clin Microbiol 40:1626-35.

66. Wise MG, Siragusa GR, Plumblee J, Healy M, Cray PJ, Seal BS. 2009. Predicting *Salmonella enterica* serotypes by repetitive sequence-based PCR. J Microbiol Methods 76:18-24.

67. Fitzgerald C, Collins M, van Duyne S, Mikoleit M, Brown T, Fields P. 2007. Multiplex, bead-based suspension array for molecular determination of common *Salmonella* serogroups. J Clin Microbiol 45:3323-34.

68. McQuiston JR, Waters RJ, Dinsmore BA, Mikoleit ML, Fields PI. 2011. Molecular determination of H antigens of *Salmonella* by use of a microsphere-based liquid array. J Clin Microbiol 49:565-73.

69. Fitzgerald C, Collins M, van Duyne S, Mikoleit M, Brown T, Fields P. 2007. Multiplex, bead-based suspension array for molecular determination of common *Salmonella* serogroups. J Clin Microbiol 45:3323-3334.

70. Köser CU, Ellington MJ, Cartwright EJ, Gillespie SH, Brown NM, Farrington M, Holden MT, Dougan G, Bentley SD, Parkhill J. 2012. Routine use of microbial whole genome sequencing in diagnostic and public health microbiology. PLoS pathog 8:e1002824.

71. Brunham LR, Hayden MR. 2012. Whole-genome sequencing: the new standard of care? Science 336:1112-1113.

72. Deng X, den Bakker HC, Hendriksen RS. 2016. Genomic Epidemiology: Whole-Genome-Sequencing–Powered Surveillance and Outbreak Investigation of Foodborne Bacterial Pathogens. Annu Rev Food Sci Technol 7:353-374.

73. Goodwin S, McPherson JD, McCombie WR. 2016. Coming of age: ten years of next-generation sequencing technologies. Nature Rev Genet 17:333-351.

74. Salipante SJ, Kawashima T, Rosenthal C, Hoogestraat DR, Cummings LA, Sengupta DJ, Harkins TT, Cookson BT, Hoffman NG. 2014. Performance comparison of Illumina and ion torrent next-generation sequencing platforms for 16S rRNA-based bacterial community profiling. Appl Environ Microbiol 80:7583-7591.

75. Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B. 2009. Real-time DNA sequencing from single polymerase molecules. Science 323:133-138.

76. Koren S, Phillippy AM. 2015. One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly. Curr Opin Microbiol 23:110-120.

77. Ku C-S, Roukos DH. 2013. From next-generation sequencing to nanopore sequencing technology: paving the way to personalized genomic medicine. Expert Rev Med Devices 10:1-6.

78. Laver T, Harrison J, O'neill P, Moore K, Farbos A, Paszkiewicz K, Studholme DJ. 2015. Assessing the performance of the Oxford Nanopore Technologies MinION. Biomol Detect Quantif 3:1-8.

79. Sović I, Šikić M, Wilm A, Fenlon SN, Chen S, Nagarajan N. 2016. Fast and sensitive mapping of nanopore sequencing reads with GraphMap. Nat Commun 7.

80. Deschamps S, Mudge J, Cameron C, Ramaraj T, Anand A, Fengler K, Hayes K, Llaca V, Jones TJ, May G. 2016. Characterization, correction and de novo assembly of an Oxford Nanopore genomic dataset from Agrobacterium tumefaciens. Sci Rep 6.

81. Wyres KL, Conway TC, Garg S, Queiroz C, Reumann M, Holt K, Rusu LI. 2014. WGS Analysis and Interpretation in Clinical and Public Health Microbiology Laboratories: What Are the Requirements and How Do Existing Tools Compare? Pathogens 3:437-58.

82. Deng X, Desai PT, den Bakker HC, Mikoleit M, Tolar B, Trees E, Hendriksen RS, Frye JG, Porwollik S, Weimer BC, Wiedmann M, Weinstock GM, Fields PI, McClelland M. 2014. Genomic epidemiology of *Salmonella enterica* serotype Enteritidis based on population structure of prevalent lineages. Emerg Infect Dis 20:1481-9.

83. Gilmour MW, Graham M, Van Domselaar G, Tyler S, Kent H, Trout-Yakel KM, Larios O, Allen V, Lee B, Nadon C. 2010. High-throughput genome sequencing of two *Listeria monocytogenes* clinical isolates during a large foodborne outbreak. BMC genomics 11:1.

84. CDC.gov. (2010, May 4). Multistate Outbreak of Human *Salmonella* Montevideo Infections (Final Update). Retrieved January 30, 2017, from https://www.cdc.gov/*Salmonella*/2010/montevideo-5-4-2010.html.

85. Lienau EK, Strain E, Wang C, Zheng J, Ottesen AR, Keys CE, Hammack TS, Musser SM, Brown EW, Allard MW, Cao G, Meng J, Stones R. 2011. Identification of a salmonellosis outbreak by means of molecular sequencing. N Engl J Med 364:981-2.

86. Grad YH, Lipsitch M, Feldgarden M, Arachchi HM, Cerqueira GC, Fitzgerald M, Godfrey P, Haas BJ, Murphy CI, Russ C, Sykes S, Walker BJ, Wortman JR, Young S, Zeng Q, Abouelleil A, Bochicchio J, Chauvin S, Desmet T, Gujja S, McCowan C, Montmayeur A, Steelman S, Frimodt-Moller J, Petersen AM, Struve C, Krogfelt KA,

Bingen E, Weill FX, Lander ES, Nusbaum C, Birren BW, Hung DT, Hanage WP. 2012. Genomic epidemiology of the *Escherichia coli* O104:H4 outbreaks in Europe, 2011. Proc Natl Acad Sci U S A 109:3065-70.

87. Beutin L, Hammerl JA, Strauch E, Reetz J, Dieckmann R, Kelner-Burgos Y, Martin A, Miko A, Strockbine NA, Lindstedt BA, Horn D, Monse H, Huettel B, Muller I, Stuber K, Reinhardt R. 2012. Spread of a distinct Stx2-encoding phage prototype among *Escherichia coli* O104:H4 strains from outbreaks in Germany, Norway, and Georgia. J Virol 86:10444-55.

88. Harris SR, Feil EJ, Holden MT, Quail MA, Nickerson EK, Chantratita N, Gardete S, Tavares A, Day N, Lindsay JA, Edgeworth JD, de Lencastre H, Parkhill J, Peacock SJ, Bentley SD. 2010. Evolution of MRSA during hospital transmission and intercontinental spread. Science 327:469-74.

89. Jackson BR, Salter M, Tarr C, Conrad A, Harvey E, Steinbock L, Saupe A, Sorenson A, Katz L, Stroika S. 2015. Notes from the field: listeriosis associated with stone fruit—United States, 2014. MMWR Morb Mortal Wkly Rep 64:282-3.

90. Cody AJ, McCarthy ND, van Rensburg MJ, Isinkaye T, Bentley SD, Parkhill J, Dingle KE, Bowler IC, Jolley KA, Maiden MC. 2013. Real-time genomic epidemiological evaluation of human Campylobacter isolates by use of whole-genome multilocus sequence typing. J Clin Microbiol 51:2526-2534.

91. Chen Y, Frazzitta AE, Litvintseva AP, Fang C, Mitchell TG, Springer DJ, Ding Y, Yuan G, Perfect JR. 2015. Next generation multilocus sequence typing (NGMLST) and the analytical software program MLSTEZ enable efficient, cost-effective, high-throughput, multilocus sequencing typing. Fungal Genet Biol 75:64-71.

92. Sheppard SK, Jolley KA, Maiden MC. 2012. A Gene-By-Gene Approach to Bacterial Population Genomics: Whole Genome MLST of Campylobacter. Genes (Basel) 3:261-77.

93. Kovanen SM, Kivisto RI, Rossi M, Schott T, Karkkainen UM, Tuuminen T, Uksila J, Rautelin H, Hanninen ML. 2014. Multilocus sequence typing (MLST) and whole-genome MLST of Campylobacter jejuni isolates from human infections in three districts during a seasonal peak in Finland. J Clin Microbiol 52:4147-54.

94. Ruppitsch W, Pietzka A, Prior K, Bletz S, Fernandez HL, Allerberger F, Harmsen D, Mellmann A. 2015. Defining and Evaluating a Core Genome Multilocus Sequence Typing Scheme for Whole-Genome Sequence-Based Typing of *Listeria monocytogenes*. J Clin Microbiol 53:2869-76.

95. Chen Y, Gonzalez-Escalona N, Hammack TS, Allard MW, Strain EA, Brown EW. 2016. Core Genome Multilocus Sequence Typing for Identification of Globally Distributed Clonal Groups and Differentiation of Outbreak Strains of *Listeria monocytogenes*. Appl Environ Microbiol 82:6258-6272.

96. Graves AK. 2011. Food safety and implications for microbial source tracking, p 585-607, Microbial Source Tracking: Methods, Applications, and Case Studies. Springer.

97. Fu LL, Li JR. 2014. Microbial source tracking: a tool for identifying sources of microbial contamination in the food chain. Crit Rev Food Sci Nutr 54:699-707.

98. Scott TM, Rose JB, Jenkins TM, Farrah SR, Lukasik J. 2002. Microbial source tracking: current methodology and future directions. Appl Environ Microbiol 68:5796-5803.

99. Scott TM, Rose JB, Jenkins TM, Farrah SR, Lukasik J. 2002. Microbial source tracking: current methodology and future directions. Appl Environ Microbiol 68:5796-803.

100. Rivera B, Rock C. 2011. Microbial source tracking: Watershed characterization and source identification. Tucson, Arizona.

101. Stoeckel DM, Harwood VJ. 2007. Performance, design, and analysis in microbial source tracking studies. Appl Environ Microbiol 73:2405-15.

102. Nayak R, Stewart-King T. 2008. Molecular epidemiological analysis and microbial source tracking of *Salmonella enterica* serovars in a preharvest turkey production environment. Foodborne Pathog Dis 5:115-26.

103. Cooley M, Carychao D, Crawford-Miksza L, Jay MT, Myers C, Rose C, Keys C, Farrar J, Mandrell RE. 2007. Incidence and tracking of *Escherichia coli* O157:H7 in a major produce production region in California. PLoS One 2:e1159.

104. Sadowsky MJ, Call DR, Santo Domingo JW. 2007. The future of microbial source tracking studies, p 235-277, Microbial Source Tracking. American Society of Microbiology.

105. Gomi R, Matsuda T, Matsui Y, Yoneda M. 2014. Fecal source tracking in water by next-generation sequencing technologies using host-specific *Escherichia coli* genetic markers. Environ Sci Technol 48:9616-9623.

106. Bergholz TM, Switt AIM, Wiedmann M. 2014. Omics approaches in food safety: fulfilling the promise? Trends Microbiol 22:275-281.

107. Khromykh A, Solomon BD. 2015. The benefits of whole-genome sequencing now and in the future. Mol Syndromol 6:108-109.

CHAPTER 3

SALMONELLA SEROTYPE DETERMINATION UTILIZING HIGH-THROUGHPUT GENOME SEQUENCING DATA [1]

Zhang S, Yin Y, Jones MB, Zhang Z, Deatherage Kaiser BL, Dinsmore BA, Fitzgerald C, Fields PI, Deng X. 2015. *J Clin Microbiol.* 53(5):1685-92.

ABSTRACT

Salmonella serotyping has been applied as the first line subtyping method in the U.S. National *Salmonella* Surveillance System. To address the challenges of traditional serotyping approaches, SeqSero (http://www.denglab.info/SeqSero), a web-based bioinformatics tool to predict *Salmonella* serotypes from whole genome sequencing (WGS) data, was developed. Based on our curated databases for the serotype determinants (*rfb* gene cluster, *wzx*, *wzy*, *fliC* and *fljB* genes, etc.), SeqSero can achieve almost full spectrum *Salmonella* serotyping (more than 2,300 serotypes). Three datasets were used to evaluate the performance of SeqSero: 1) raw sequencing reads from 308 *Salmonella* isolates of known serotypes which confirmed by Centers for Disease Control and Prevention (CDC); 2) raw sequencing reads from 3,306 *Salmonella* isolates sequenced and uploaded to public available databases by GenomeTrakr, a U.S. national monitoring network operated by the Food and Drug Administration; and 3) 354 publicly available draft or complete *Salmonella* genomes from NCBI Genbank database. The evaluation showed that SeqSero can reliably predict serotypes from WGS data with a high accuracy and fast speed. In addition, SeqSero detected signal of serotype Typhimurium in metagenome data from mice orally challenged with *Salmonella enterica* serovar Typhimurium strain 14028s.

Key words: *Salmonella* serotyping, SeqSero, whole genome sequencing

1. Introduction

Salmonella is one of the most common causes of food poisoning in the United States (1). This pathogen can spread by various food vehicles such as poultry, eggs, beef, and milk products, etc. (2). Since 1960s, the U.S. National *Salmonella* Surveillance System has been built on *Salmonella* serotyping, which is a subtyping method traditionally performed based on the immunologic agglutination between *Salmonella* ~~cell~~ antigens, i.e. lipopolysaccharide O antigen and flagellar H antigens, and their corresponding antisera. Based on combinations of different types of O and H antigens, *Salmonella* can be subtyped into different serotypes. So far, according to the White-Kauffmann-Le minor scheme (3), more than 2,500 *Salmonella* serotypes have been recorded.

Although still regarded as vital in *Salmonella* serotyping, the traditional phenotype-based approach has multiple shortcomings, such as being labor-intensive, time-consuming, and logistically challenging in preparation of hundreds of antisera for full-set serotype identification. Thus molecular *Salmonella* serotyping methods have been developed based on two strategies. First, other subtype schemes that display correlation with certain serotypes have been used to predict these serotypes. For example, PFGE patterns (4, 5), MLST profiles (6) and repetitive sequence-based PCR patterns (7) have been applied for *Salmonella* serotype prediction. However, the accuracy of such approaches can be compromised if horizontal gene transfer of either the surrogate markers or the genetic determinants of serotype occurs. In addition, these indirect methods require additional validation on the association between surrogate markers and corresponding serotypes. Second, genetic determinants of serotypes i.e. *rfb* gene cluster (8, 9), *fliC (10)* and *fljB* (11) genes have been directly targeted for serotype determination. For example, a Bio-Plex platform based assay for the detection of common O antigens and H antigens has been developed for genetic subtyping of *Salmonella* (12, 13). This type of method has the

advantage of maintaining the important continuity between the phenotypic and genotypic determination of serotypes (14, 15).

WGS has emerged as a powerful tool for public health microbiology (16) and been increasingly applied in foodborne pathogen surveillance and outbreak investigation (17). The routine sequencing of microbial pathogen genomes has created a large amount of WGS data in publicly available databases, such as GenomeTrakr (NCBI Bioproject#: 183844), NCBI Genbank, etc. A promising benefit of WGS is to combine multiple subtyping and pathogen characterization methods into a single WGS platform. Efforts have already been made along that direction. For example, studies have been conducted to predict multi-locus sequence typing (18, 19), determine antimicrobial resistance genes (20) and identify virulence profiles (21) based on WGS data of foodborne pathogens.

In this study, a bioinformatics tool named SeqSero was developed to predict *Salmonella* serotypes by identifying the types of major genetic determinants of *Salmonella* serotypes (*rfb* gene cluster, *wzx*, *wzy*, *fliC* and *fljB* gene) from WGS data. By evaluating the performance through CDC genomes with confirmed serotypes and publicly available genomes with annotated serotypes, we demonstrated that SeqSero allowed robust, fast and comprehensive prediction of *Salmonella* serotype for both raw sequencing reads and genome assembly data.

2. Materials and Methods

2.1 Strains and genomes

A total of three different datasets of genomes were used to evaluate the performance of SeqSero.

First, a total of 308 CDC isolates of 72 experimentally confirmed serotypes were collected (Supplementary Table S3.1). Among them, 229 were from relatively uncommon serotypes and

sequenced on an Illumina HiSeq instrument (100 bp, paired end reads) following manufacturer's instructions. The other 79 genomes WGS raw sequencing reads genomes were from common *Salmonella* serotypes archived in the 100K Food Pathogen Project (NCBI BioProject PRJNA186441). The serotypes of the isolates were confirmed through both phenotypic and genetic (12, 13) serotyping techniques by CDC.

Second, genomes from GenomeTrakr depository of Food and Drug Administration (NCBI BioProject 183844) as of June 1st 2014 were downloaded and analyzed, excluding the genomes: i) with no serotype or more than one serotypes annotated (n = 766); ii) annotated as rough, nonmotile strains (n = 39); iii) annotated as monophasic variants (n = 76); and iv) with sequencing coverage less than 10 (n = 11). Overall, a total of 3,306 isolates with 228 annotated serotypes in raw sequencing reads format were included for the evaluation analysis (Supplementary Table S3.2).

Third, the draft or complete genomes of *Salmonella* isolates from NCBI GenBank as of April 1st, 2014 were downloaded and analyzed. The genomes without available serotype information or whose draft assemblies had a N50 lower than 150,000 bp were excluded. In total, 354 draft or complete genomes with 44 annotated serotypes were included in this dataset (Supplementary Table S3.3).

2.2 *Salmonella* serotype determinant database

Three primary databases were built for serotype determinants: i) a database of the 46 described *rfb* clusters (22) for O antigen prediction from genome assemblies; ii) a database containing *wzx* (O-antigen flippase), *wzy* (O-antigen polymerase) and other genomic targets (see details below) for O antigen prediction from raw sequencing reads; ii) a database of various types of *fliC* and *fljB* gene alleles primarily from a previous study (13) for H antigen prediction.

All the serotype determinants were downloaded from NCBI GenBank or extracted from publicly available *Salmonella* genomes. They were manually curated before adding to the databases.

Additional markers were used to distinguish two O antigen groups that have high levels of sequence similarity: the O9 group that contains O9, O2, O9,46, and O9,46,27, and the O3 group that contains O3,10 and O1,3,19 (Table 3.1). For example, a frameshift mutation in *tyv* gene was used to differentiate O2 and O9 (23). Also, a small *rfb* sequence specific to serotype O3,10 was utilized to confirm the presence of O3,10 or O1,3,19. All the genomic markers were evaluated by more than 30 genomes of each antigen type (data not shown).

Table 3.1 Additional markers for differentiating O groups O2, O9, O9,46, O9,46,27, O3,10 and O1,3,19

Marker	Distribution profile [a]					
	O9	O2	O9,46	O9,46,27	O3,10	O1,3,19
wzx (O9)	+	+	+	+	-	-
wzy (O9,46)	-	-	+	-	+	+
wzy (O9,46,27)	-	-	-	+	-	-
wzy (O3,10)	-	-	-	-	+	+
rfb sequnence specific to O3, 10	-	-	-	-	+	-
SNP in *tyv* [b]	-	+	-	-	-	-

[a] "+" and "-" reprents the presence and absence of the marker in a particular O antigen. Note that each O antigen listed here features distinct combined distribution profile for the six markers.
[b] A frame shift mutation in the gene (23)

For the two O antigen databases, each of the 46 O antigens was represented by a single *rfb* cluster (22) or a single allele of the *wzx* or *wzy* gene (24). However, for H antigen databases, to address the challenge caused by the high similarity among some specific H antigens, multiple but distinct alleles for same H antigen type were present in the databases. Thus, for such situation, a multiple rounds based reads mapping approach (through Burrows-Wheeler Aligner, BWA (25))

was developed (see details below), which required three additional datasets or sub-databases of H antigen sequences. (i) The clustering scheme to describe the groups formed by *fliC* and *fljB* alleles based on their sequence similarity (Supplementary Table S3.4). This clustering was utilized to determine the most possible H antigen group after the first two rounds of reads mapping (see details below). (ii) A database consisted of representative alleles of each H antigen type in the most possible H antigen group (one allele for each type) to extract mapped reads in the third round of reads mapping. The alleles were selected following the standard that it was near the midpoint of the phylogenetic tree of all the alleles for an antigen type to ensure the representativity. (iii) A database containing the middle, variable sequences of all the alleles for every antigen in the most possible H antigen group, which was applied to make BLAST analysis between the database and the extracted reads to confirm the final H antigen type.

All the aforementioned databases and additional datasets have been well curated and regularly updated at www.denglab.info/SeqSero.

2.3 SeqSero pipeline

The major components and workflows of the SeqSero system were outlined in Figure 3.1. Specifically, based on the curated databases of *Salmonella* serotyping determinants, combined with the application of bioinformatics tools such as BLAST and BWA, two major workflows including serotype determination from (i) genome assembly and (ii) raw sequencing reads, were described (see details below).

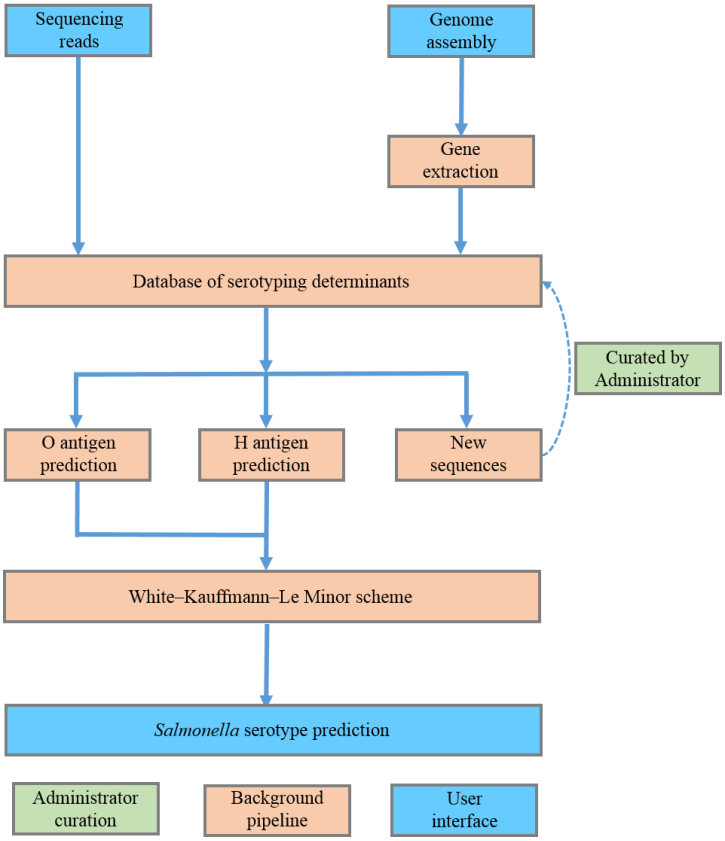

Figure 3.1. A schematic overview of the SeqSero Pipeline.
Two workflows are represented, including serotype determination from (i) genome assembly and (ii) raw sequencing reads.

2.4 Serotype antigen prediction from genome assembly

For O antigen prediction, the *rfb* gene cluster was firstly located by aligning its two flanking genes, i.e. *galF* and *gnd* genes, against the target genome assembly. If the two genes were in same contig, the sequence between the two loci was extracted, i.e. the complete sequence of *rfb* gene cluster. If the two genes were not in same contig, the four contig fragments flanking sequences of *galF* and *gnd* genes were extracted, i.e. the potential partial sequences of *rfb* gene cluster. The extracted sequences of target genome were compared with the *rfb* database using BLAST. The aligned *rfb* allele with highest score was selected as the most possible prediction for O antigen.

For H antigen prediction, the *fliC* and *fljB* sequences of the target genome were extracted by *in silico* PCR (http://hgwdev.cse.ucsc.edu/~kent/src/) using primers in Table 3.2. Due to the variability of flanking regions of *fljB* genes, multiple sets of *fljB* primers were applied. The extracted *fliC* and *fljB* genes were aligned to H antigen database using same method as O-antigen prediction through BLAST.

Table 3.2 Primers for *in silico* PCR

Target gene	Designation	Sequence
fliC	F15/F14s (26)	GAAATTCAGGTGCCGATACAAGGG **&** CGCTGCCTTGATTGTGT
fljB	sense 56/Flj4 R (26)	TGTCGATAACCTGGATGACACAGG **&** GGCATCCAGTGTAGTACCATTATC
fljB	F/R (26)	TCGATAACCTGGATGACACAG **&** CATTTACAGCTATACATTCCATAAAGA
fljB	*fljB* atgfor/*fljB* stoprev (27)	CCGAATTCATGGCACAAGTAATCAACACTAA **&** CGGGATCCTTAACGTAACACAGACAGCACG
fljB	MR-22 *fljB*F/MR-23 *fljB*R (27)	GGCACAAGTAATCAACACTAACA **&** CATTTACAGCCATACATTCCATA

2.5 Serotype antigen prediction from raw sequencing reads

For O antigen prediction, a straightforward reads mapping-based strategy was developed. In brief, the raw sequencing reads were firstly mapped to the database containing *wzx/wzy* and other serotype-specific genome markers using BWA in default parameter setting. Then the allele with highest number of mapped reads was chosen as the prediction for O antigen.

For H antigen prediction, the same method applied in O antigen prediction was not suitable due to the high similarity of some *fliC* and *fljB* alleles and the redundant nature of H antigen databases. Specifically, when multiple, closely related alleles were present in the database, the sequencing reads which should be mapped to the single and correct allele, may be diluted by

multiple high similarity alleles, diminishing the number of reads mapped to the correct allele, which increased the prediction error for H antigen (Figure 3.2, called “dilution effect”). To address the challenge, a two tier step approach based on BWA mapping and BLAST analysis was developed. Briefly, the first step was to apply the first two rounds of BWA mapping to confirm the most possible group (described in Supplementary Table S3.4). If multiple antigenic types existed in the H antigen group, the second step was to utilize the third round mapping to the representative alleles in the most possible group to extract the mapped reads, then BLAST analysis was made between the extracted reads and the middle, variable regions of the alleles of the most possible group. The allele ranked highest in BLAST score was the most possible H-antigen prediction.

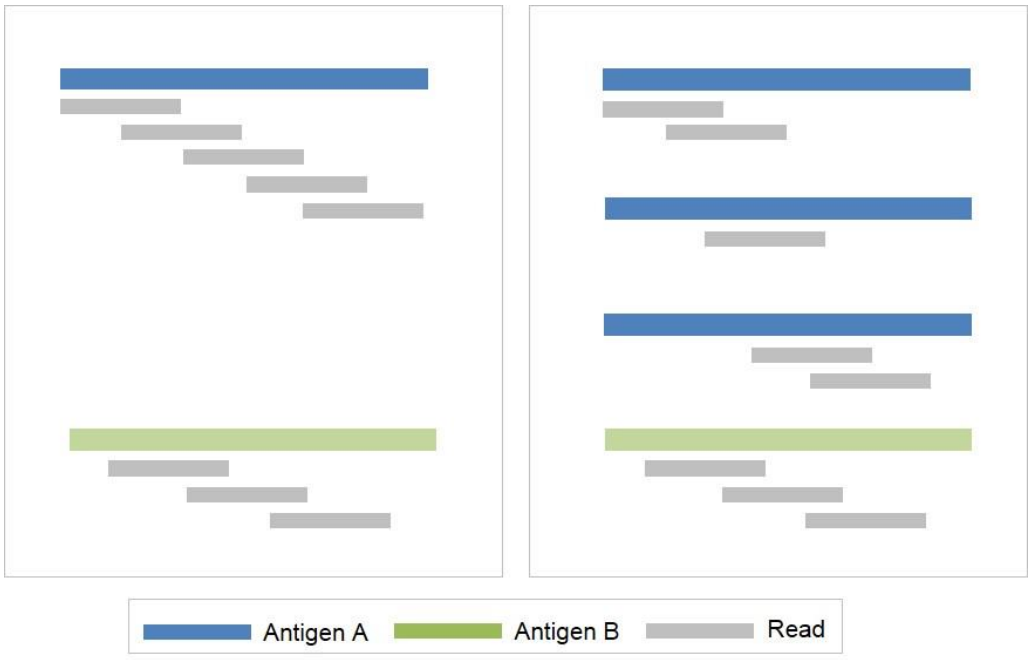

Figure 3.2. The reads dilution effect.
When multiple alleles of Antigen A (right) are archived in the database instead of only one allele (left), those alleles compete for reads, leading to a decrease of the highest amount of reads mapped to an Antigen A allele from 5 to 2, fewer than the 3 mapped to Antigen B.

Here, we applied an example, the raw sequencing reads of a serotype Typhimurium genome (NCBI SRA accession no. SRX528051), to illustrate how SeqSero called *fliC* type from raw sequencing reads (Figure 3.3). The first mapping of raw reads to H antigen database confirmed the top three antigen groups with highest number of mapped reads were *fliC* "e,h" > "i,r" > "z35". This order was not the correct one due to the presence of "dilution effect" error (Figure 3.2). So the representative alleles of the top 3 antigens were selected and formed a temporary, small and non-redundant database containing only the 3 representative alleles. The second mapping of raw sequencing reads to the temporary database corrected the "dilution effect" error (Figure 3.2) and showed the correct order was *fliC* "i,r" > "e,h" > "z35" based on the number of mapped reads. The next step was to confirm the most possible antigen type in the most possible antigen group, i.e. to predict the antigen type in group *fliC* "i,r". Thus, the third mapping of raw sequencing reads to the database containing representative alleles of the antigenic types in group *fliC* "ir" were employed to extract relevant reads with homology to its antigenic types. At last, the BLAST analysis was applied to align the reads to the variable regions of all the alleles of *fliC* "ir" group. The highest score pointed to the most likely allele and its corresponding antigen was "i". By similar way, the *fljB* antigen was also predicted to be "1,2".

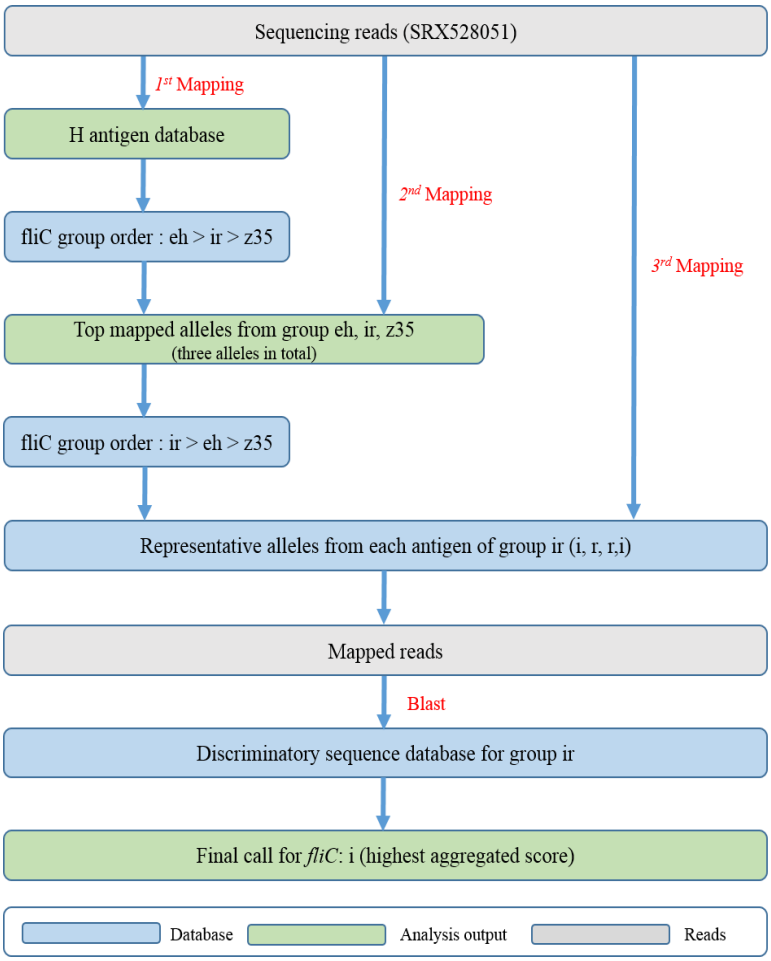

Figure 3.3. An example of the two-tier workflow of *fliC* allelic type prediction using a Typhimurium (4:i:1,2) raw sequencing genome .
The details of predefined antigen groups were present in Supplementary Table S3.4.

2.6 Statistical analysis

The ability of using the top allele with largest number of mapped reads to predict H antigen type was evaluated by GenomeTrakr dataset. Specifically, we aligned the raw sequencing reads to the H antigen databases (both *fliC* and *fljB*). Then for each raw sequencing reads genome, a scaled reads difference (α) was calculated by following formula:

$$\alpha = [(x - y)/z] \times 10^6.$$

where x is the number of aligned reads to the top allele, y is the number of aligned reads to the second top allele, z is the number of all sequencing reads which applied to scale the difference since the larger number of sequencing reads (z) tends to create a bigger reads difference. We utilized logistic regression to estimate the probability of making an incorrect identification given the scaled size of the mapped reads difference α. The outcome of the model was a binary indicator of whether the correct H antigen was determined.

2.7 Metagenomics data and phylogenetic analysis

The metagenomics data was achieved from one previous study (28). The metagenomics DNA sequencing data samples from day -1 (NCBI accession# SRR916930), day 3 (NCBI accession# SRR916932), day 6 (NCBI accession# SRR916933), day 14 (NCBI accession# SRR916931) were downloaded. SeqSero was used to detect serotype signals from the samples with the same method for pure culture WGS data.

2.8 Phylogenetic analysis on metagenomics data

Raw reads from each metagenome sample were mapped to the complete genome (GenBank accession number CP001363) of the strain str. 14028s through BWA in default settings. The high quality single nucleotide polymorphisms (SNPs) were identified and a core genome SNPs based maximum likelihood tree was built using similar methods as described in our previous study (29).

2.9 Nucleotide sequence accession numbers

The WGS data which were sequenced in this study have been deposited in the NCBI Sequence Read Archive under sample accession numbers SAMN03264859 to SAMN03264906, SAMN03264909 to SAMN03265006, and SAMN03265010 to SAMN03265087.

3. Results and Discussion

3.1 Databases of antigen determinants

For H antigen database, there were a total of 473 alleles representing 56 antigenic types for *fliC* genes and a total of 190 alleles representing 18 antigenic types for *fljB* genes. For O antigen database, firstly, 46 *rfb* gene clusters were included in the *rfb* database which was used to predict O antigen type from genome assemblies. Secondly, a database consisted of *wzx*, *wzy*, and other genomic markers (Table 3.2) covering all 46 O antigen types were applied for raw sequencing reads O antigen prediction. Theoretically, our databases can allow 2,389 of the 2,577 serotypes described in the White-Kauffmann-Le Minor scheme.

3.2 Robustness of H antigen identification by reads mapping

The inclusion of a set of redundancy existed, highly related but phenotypically different alleles in H antigen database, such as the alleles in G complex and 1 complex (details in Supplementary Table S3.4), constituted a challenge to do robust identification for H antigen type. Thus we evaluated the ability of our reads mapping approach by calculating the scaled reads difference (α) for each genome of GenomeTrakr dataset. The logistically regression model displayed that the median scaled reads difference values were 3.59 for *fliC* and 1.82 for *fljB*, corresponding to predicted probabilities of an incorrect antigen call of 2.7% and 5.6% (Figure 3.4), which displayed the robustness of our reads mapping based approach. It was noted that it was just the result based on the first round of mapping (Figure 3.3), therefore, the accuracy and robustness would be expected to increase with the combination of subsequent mapping and BLAST analyses to address the challenges such as “dilution effect” (Figure 3.2).

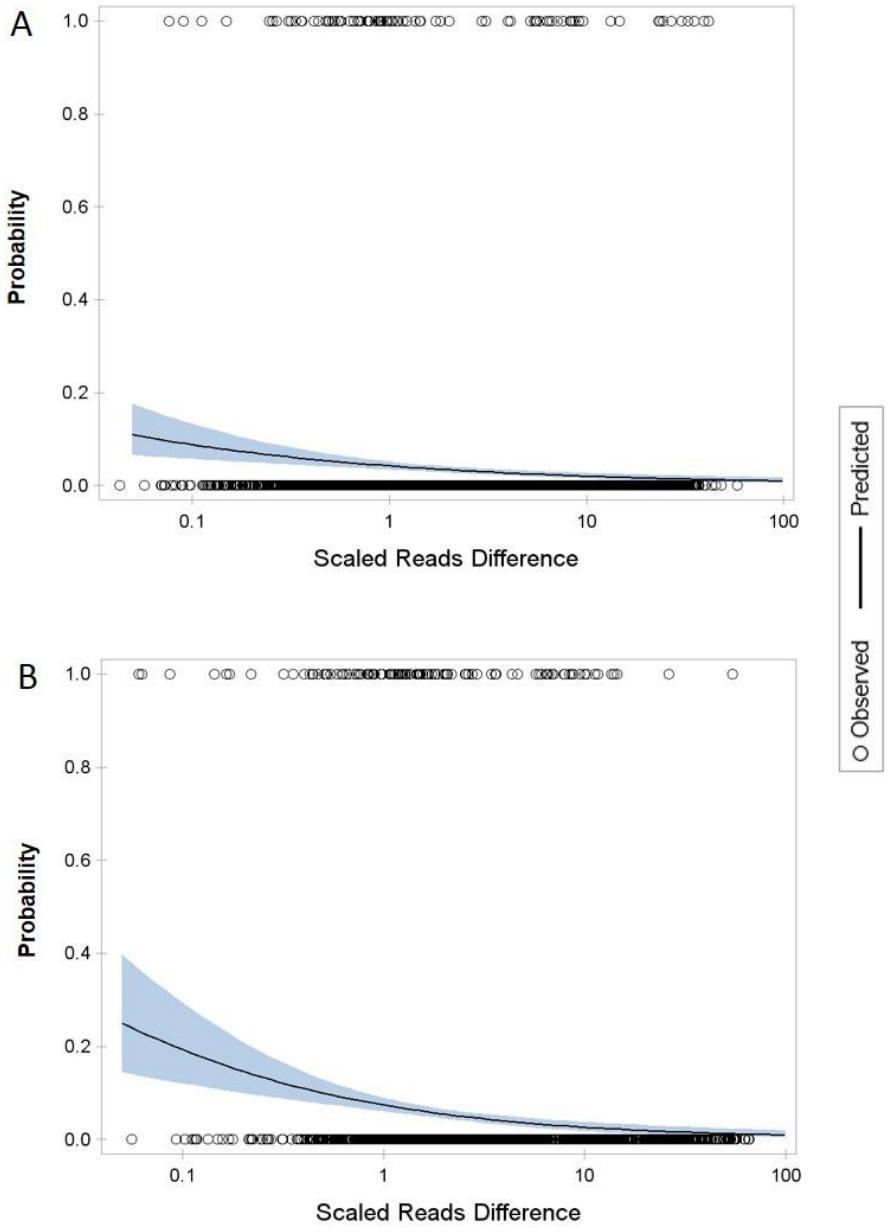

Figure 3.4. The logistic regression for predicted incorrect H antigen identification using reads mapping approach with 95% confidence interval. A: Prediction for *fliC* identification. B: Prediction for *fljB* identification.

Logistic regression was used to estimate the probability of making an incorrect identification given the size of scaled reads difference (α) on GenomeTrakr dataset. All the data was based on the first round of reads mapping through BWA.

3.3 Serotype prediction from pure culture WGS data

We summarized the prediction results of the three previously described datasets used for SeqSero evaluation (details in Materials and Methods section) in Table 3.3. The CDC dataset and GenomeTrakr dataset were aimed to evaluate the performance of SeqSero serotyping for raw sequencing reads. The NCBI assembled genome dataset was applied to evaluate the performance of SeqSero serotyping for genome assemblies.

Table 3.3. Accuracy of serotype predictions.

Result	Number of genomes (% of total)		
	Reads mapping, CDC dataset	Reads mapping, GenomeTrakr dataset	Assembled genome dataset
Expected Serotype [a]	304 (98.7%)	3,061 (92.6%)	324 (91.5%)
Unexpected Serotype [b]	2 (0.65%)	205 (6.2%)	11 (3.1%)
Partial or no Serotype [c]	2 (0.65%)	40 (1.2%)	19 (5.4%)
Total tested	308	3306	354

[a] Predicted serotype was considered correct if predicted serotype antigens corresponded to antigens detected by conventional methods. For GenomeTrakr and genome assembly datasets, serotype predictions in consensus with annotated serotype were considered correct.

[b] Numbers represent serotype prediction inconsistent with annotated serotype; the accuracy of the annotated serotype is unknown.

[c] One or more than one of the expected serotype antigen were not detected.

For the confirmed CDC dataset with 308 genomes included, 304 isolates (98.7%) were identified to have the same serotypes as confirmed, 2 isolates were predicted as containing partial serotype information, i.e. one or more than one antigens showed no prediction results, and 2 other isolates were showed unexpected results with confirmed serotypes. For the other two annotated datasets, i.e. GenomeTrakr dataset and assembled genome dataset, the accuracies were 92.6% and 91.5% respectively.

To evaluate whether the four genomes without producing "expected serotype" in the CDC confirmed dataset caused by the problem of SeqSero or by other factors, further analyses were made. The two Hvittingfoss (antigenic formula I 16:b:e,n,x) genomes producing no O antigen in the predictions were showed lacking sequencing reads that mapped to any *rfb* cluster that included its *wzx* gene and the *wzy* gene. Since O antigen was confirmed by CDC in these strains using conventional methods, it indicated the incompleteness of sequencing to cover the whole *rfb* gene cluster region. For the genome of London (antigenic formula I 3,10:l,v:1,6) of which the *fljB* type was wrongly predicted to be "e,n,x", the sequencing reads that can be assembled into both *fljB* "1,6" and "e,n,x" alleles were found. In addition, we found two possible *fljB* allele sequences in its genome assembly. This implied the strain may contain a third flagellin allele, of which the phenomenon has been described before (30). The similar situation also occurred for one Weltevreden (antigenic formula I 3,10:r:z6) genome of which *fliC* type was wrongly predicted to be "i", and reads related to both *fliC* "r" and "i" were found in the WGS data.

The raw sequencing reads format is more preferred than genome assembly for SeqSero. This was supported by the observation that the accuracy of assembled genome dataset was the lowest in all three datasets (Table 3.3). Because firstly, the failure in extracting serotype determinants may occur for draft genome assemblies. The determinant sequences, i.e. *rfb* gene cluster, *fliC* and fliB genes, might not locate in the same assembly contig of the genome, causing the failure of extracting complete determinant sequences which would lower the accuracy. Secondly, determining serotypes form raw sequencing reads directly allows users to save the time and platforms to do high quality genome assembly. Based on the test with Genometrakr dataset, in average it would take less than 5 min for a Linux based computational platform, with single

2.50GHz CPU core and 4 GB of random access memory (RAM), to identify serotype from one raw sequencing reads genome (the dataset has an average of 2.17 million reads per genome).

The major challenge for reads mapping approach is how to accurately call H antigen type from the redundancy-existed database containing multiple highly related but phenotypically different alleles. Thus a two tier step approach based on combination of BWA reads mapping for efficiency and BLAST analysis for resolution was developed (Figure 3.2). In general, the first two rounds of reads mapping were applied to confirm the most possible H antigen group (Supplementary Table S3.4). If the group contained two or more than two antigen types, then a third round mapping was used to extract the reads that could be aligned to the correct H antigen genes. A following BLAST analysis between the aligned reads and the alleles of that antigen group was made to identify the most possible allele. The potential of this approach was demonstrated by the high accuracy of the two raw sequencing reads datasets.

In the two datasets of raw sequencing reads, GenomeTrakr dataset was somewhat lower than CDC dataset in accuracy. Two reasons may explain. First, since the strains of GenomeTrakr were serotyped and sequenced by a variety of laboratories, it is possible that some genomes with wrong serotyping information or mislabeling were uploaded to GenomeTrakr without Quality Assurance. We checked the genomes without “expected serotype” results by SeqSero, and the genome analyses of most of them pointed to the problem of mislabeling and low quality sequencing (details in Supplementary Table S3.2). However, we were unable to confirm them since the strains were not available to us. Second, the GenomeTrakr dataset is more diverse and larger than CDC dataset, thus some of the antigen sequences may not characterized and included in our curated database yet, which may cause “unexpected serotype” or “partial serotype”

predictions. This also demonstrated that it is important to keep updating and curating our databases when new alleles are available.

On the whole, SeqSero provided accurate predictions for *Salmonella* serotyping. In the three evaluation datasets, even the lowest accuracy was 91.5%, which was for genome assembly dataset with extraction failure issue described previously. The raw sequencing reads datasets achieved better accuracy. In addition, based on the results of all three datasets, a total of 200 serotypes were successfully predicted (details in Supplementary Table S3.5), which covered 85 serotypes in the top 100 most commonly reported, clinical *Salmonella* serotype in the U.S. national *Salmonella* surveillance system (31).

3.4 Serotype prediction from metagenomics data

SeqSero detected Typhimurium serotype signal from all the metagenomics samples of raw sequencing reads (Table 3.4). Out of our expectation, the sample "day -1" (one day before oral challenge) contained weak Typhimurium serotype signal. But the number of mapped reads was far fewer than other samples (only four reads mapped to H antigen), and it was phylogenetically distinct from the *S. enterica* serotype Typhimurium strain 14028s compared with other samples (Figure 3.5). It indicated the sequencing reads were not likely from the strain applied in oral challenge but from strains with other origins. Even though more datasets need to be used to evaluate its ability, at least it demonstrated the potential of SeqSero in detecting *Salmonella* serotypes from metagenome sequences, which might be used in predicting the *Salmonella* pathogen serotypes for culture-independent diagnosis from fecal samples.

Table 3.4. Serotype determination from stool metagenomes of mice orally infected with *Salmonella.*

Sample		Number of reads mapped to individual antigen alleles [c]		
Sampling time	Accession [a]	*wzx/wzy* (O4) [b]	*fliC* (i)	*fljB* (1,2)
Day -1	SRR916930	273	2	2
Day 3	SRR916932	521	10	11
Day 6	SRR916933	519	12	10
Day 14	SRR916931	1572	21	21

[a] NCBI SRA accession number of the metagenome sequence.
[b] Predicted antigen type.
[c] The number of reads aligned to the best mapped antigen allele after the first round of reads mapping.

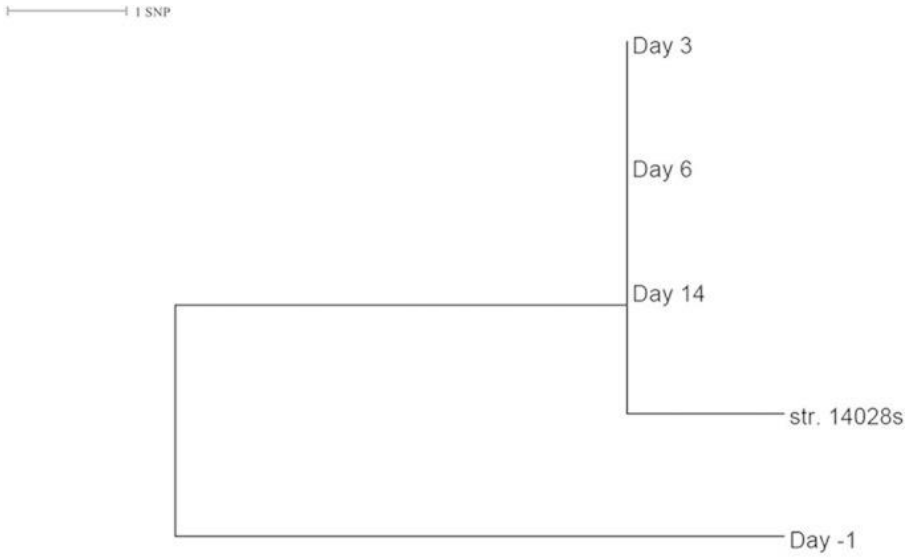

Figure 3.5. Phylogenetic relationship among detected *Salmonella enterica* serotype Typhimurium strains from mice fecal metagenomes.
Raw sequencing reads from each metagenome sample were mapped to the complete genome of strain str. 14028s (NCBI accession# CP001363). The high quality SNPs were identified and a core genome SNPs based maximum likelihood tree was built using similar methods as described in one previous study (29).

4. Conclusion

SeqSero can achieve comprehensive, accurate and fast prediction for *Salmonella* serotypes using raw WGS reads and draft genome assemblies. In addition, initial testing using mouse fecal

metagenomes suggested that SeqSero has the potential for culture-independent *Salmonella* serotyping from noise-rich metagenome data. Overall, SeqSero can provide an alternative in *Salmonella* serotyping, which would help to bridge the gap between the well-established utility of *Salmonella* serotyping and the increasing application of WGS.

References:

1. Bhunia A. 2007. Foodborne microbial pathogens: mechanisms and pathogenesis. Springer Science & Business Media.
2. Foodsafety.gov. (n.d.) *Salmonella* [Web log post]. Retrieved January 04, 2017, from https://www.foodsafety.gov/poisoning/causes/bacteriaviruses/*Salmonella*/index.html.
3. Grimont PA, Weill FX. 2007. Antigenic formulae of the *Salmonella* serovars. WHO collaborating centre for reference and research on *Salmonella*.
4. Zou W, Lin WJ, Foley SL, Chen CH, Nayak R, Chen JJ. 2010. Evaluation of pulsed-field gel electrophoresis profiles for identification of *Salmonella* serotypes. J Clin Microbiol 48:3122-3126.
5. Zou W, Lin WJ, Hise KB, Chen HC, Keys C, Chen JJ. 2012. Prediction system for rapid identification of *Salmonella* serotypes based on pulsed-field gel electrophoresis fingerprints. J Clin Microbiol 50:1524-32.
6. Kotetishvili M, Stine OC, Kreger A, Morris JG, Jr., Sulakvelidze A. 2002. Multilocus sequence typing for characterization of clinical and environmental *Salmonella* strains. J Clin Microbiol 40:1626-35.
7. Wise MG, Siragusa GR, Plumblee J, Healy M, Cray PJ, Seal BS. 2009. Predicting *Salmonella enterica* serotypes by repetitive sequence-based PCR. J Microbiol Methods 76:18-24.
8. Samuel G, Reeves P. 2003. Biosynthesis of O-antigens: genes and pathways involved in nucleotide sugar precursor synthesis and O-antigen assembly. Carbohydr Res 338:2503-19.

9. Jiang XM, Neal B, Santiago F, Lee SJ, Romana LK, Reeves PR. 1991. Structure and sequence of the *rfb* (O antigen) gene cluster of *Salmonella* serovar typhimurium (strain LT2). Mol Microbiol 5:695-713.

10. Smith NH, Selander RK. 1990. Sequence invariance of the antigen-coding central region of the phase 1 flagellar filament gene (*fliC*) among strains of *Salmonella* typhimurium. J Bacteriol 172:603-9.

11. Vanegas RA, Joys TM. 1995. Molecular analyses of the phase-2 antigen complex 1,2,.. of *Salmonella* spp. J Bacteriol 177:3863-4.

12. Fitzgerald C, Collins M, van Duyne S, Mikoleit M, Brown T, Fields P. 2007. Multiplex, bead-based suspension array for molecular determination of common *Salmonella* serogroups. J Clin Microbiol 45:3323-34.

13. McQuiston JR, Waters RJ, Dinsmore BA, Mikoleit ML, Fields PI. 2011. Molecular determination of H antigens of *Salmonella* by use of a microsphere-based liquid array. J Clin Microbiol 49:565-73.

14. Fitzgerald C, Collins M, van Duyne S, Mikoleit M, Brown T, Fields P. 2007. Multiplex, bead-based suspension array for molecular determination of common *Salmonella* serogroups. J Clin Microbiol 45:3323-3334.

15. McQuiston JR, Waters RJ, Dinsmore BA, Mikoleit ML, Fields PI. 2011. Molecular determination of H antigens of *Salmonella* by use of a microsphere-based liquid array. J Clin Microbiol 49:565-573.

16. Kwong JC, McCallum N, Sintchenko V, Howden BP. 2015. Whole genome sequencing in clinical and public health microbiology. Pathology 47:199-210.

17. McGann P, Bunin JL, Snesrud E, Singh S, Maybank R, Ong AC, Kwak YI, Seronello S, Clifford RJ, Hinkle M, Yamada S, Barnhill J, Lesho E. 2016. Real time application of whole genome sequencing for outbreak investigation - What is an achievable turnaround time? Diagn Microbiol Infect Dis 85:277-82.

18. Larsen MV, Cosentino S, Rasmussen S, Friis C, Hasman H, Marvig RL, Jelsbak L, Sicheritz-Ponten T, Ussery DW, Aarestrup FM, Lund O. 2012. Multilocus sequence typing of total-genome-sequenced bacteria. J Clin Microbiol 50:1355-61.

19. Inouye M, Conway TC, Zobel J, Holt KE. 2012. Short read sequence typing (SRST): multi-locus sequence types from short reads. BMC Genomics 13:338.

20. Zankari E, Hasman H, Cosentino S, Vestergaard M, Rasmussen S, Lund O, Aarestrup FM, Larsen MV. 2012. Identification of acquired antimicrobial resistance genes. J Antimicrob Chemother 67:2640-4.

21. Joensen KG, Scheutz F, Lund O, Hasman H, Kaas RS, Nielsen EM, Aarestrup FM. 2014. Real-time whole-genome sequencing for routine typing, surveillance, and outbreak detection of verotoxigenic *Escherichia coli*. J Clin Microbiol 52:1501-10.

22. Liu B, Knirel YA, Feng L, Perepelov AV, Senchenkova SN, Reeves PR, Wang L. 2014. Structural diversity in *Salmonella* O antigens and its genetic basis. FEMS Microbiol Rev 38:56-89.

23. Verma NK, Quigley NB, Reeves PR. 1988. O-antigen variation in *Salmonella* spp.: *rfb* gene clusters of three strains. J Bacteriol 170:103-107.

24. Fitzgerald C, Sherwood R, Gheesling LL, Brenner FW, Fields PI. 2003. Molecular analysis of the *rfb* O antigen gene cluster of *Salmonella enterica* serogroup O:6,14 and development of a serogroup-specific PCR assay. Appl Environ Microbiol 69:6099-105.

25. Li H, Durbin R. 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 26:589-95.

26. McQuiston J, Parrenas R, Ortiz-Rivera M, Gheesling L, Brenner F, Fields P. 2004. Sequencing and comparative analysis of flagellin genes *fliC*, *fljB*, and *flpA* from *Salmonella*. J Clin Microbiol 42:1923-1932.

27. Ranieri ML, Shi C, Moreno Switt AI, den Bakker HC, Wiedmann M. 2013. Comparison of typing methods with a new procedure based on sequence characterization for *Salmonella* serovar prediction. J Clin Microbiol 51:1786-97.

28. Deatherage Kaiser BL, Li J, Sanford JA, Kim YM, Kronewitter SR, Jones MB, Peterson CT, Peterson SN, Frank BC, Purvine SO, Brown JN, Metz TO, Smith RD, Heffron F, Adkins JN. 2013. A Multi-Omic View of Host-Pathogen-Commensal Interplay in *Salmonella*-Mediated Intestinal Infection. PLoS One 8:e67155.

29. Deng X, Desai PT, den Bakker HC, Mikoleit M, Tolar B, Trees E, Hendriksen RS, Frye JG, Porwollik S, Weimer BC, Wiedmann M, Weinstock GM, Fields PI, McClelland M. 2014. Genomic epidemiology of *Salmonella enterica* serotype Enteritidis based on population structure of prevalent lineages. Emerg Infect Dis 20:1481-9.

30. Smith NH, Selander RK. 1991. Molecular genetic basis for complex flagellar antigen expression in a triphasic serovar of *Salmonella*. Proc Natl Acad Sci U S A 88:956-60.

31. CDC.gov. (n.d.) National Enteric Disease Surveillance: *Salmonella* Annual Report, 2013. Retrieved January 20, 2017, from https://www.cdc.gov/nationalsurveillance/pdfs/*Salmonella*-annual-report-2013-508c.pdf

CHAPTER 4

POPULATION STRUCTURE AND SOURCE-ASSOCIATED LINEAGES OF *SALMONELLA ENTERICA* SEROTYPE TYPHIMURIUM [2]

[2] Zhang S and Deng X. To be submitted to *Nature microbiology*.

ABSTRACT:

Salmonella enterica serovar Typhimurium (ST) is a broad-host-range pathogen which can infect humans mainly through the consumption of contaminated food. A total of 1,267 ST genomes from clinical and various animal, food and environmental sources were analyzed to identify population groups (major phylogenetic lineages) and clades (smaller phylogenetic groups within a group) as well as their association with particular sources. A phylogenetic tree was constructed using whole genome sequencing SNPs (wgSNPs). A total of 10 major population groups were identified. Clustering of isolates from the same source was observed in 6 population groups, including clusters overrepresented by isolates from poultry, bovine, swine, and wild birds. Analyses of evolutionary relationship, metabolic profile, gene contents and pseudogene distribution provide further support for the source-cluster association. In addition, the potential of WGS-based subtyping in microbial source tracking (MST) for foodborne pathogen was demonstrated based on the inferred population structure of ST.

Key words: *Salmonella enterica* serovar Typhimurium, whole genome sequencing, pseudogene, microbial source tracking, population structure

1. Introduction

Salmonella enterica Serotype Typhimurium (ST) has been recorded as one of the most frequently isolated serotypes in the United States (1). According to surveillance data, some *Salmonella* serotypes are highly associated with or even restricted to specific host animals, for example, *Salmonella enterica* serotype Dublin to cattle, *Salmonella enterica* serotypes Typhi and Paratyphi A to human and *Salmonella enterica* serotype Choleraesuis to swine (2). However, there are also many generalist *Salmonella* serotypes that are associated with a wide range of hosts and sources. ST is one of the serotypes with the most diverse environmental and host distribution in *Salmonella* (3, 4). Although regarded as a broad-host-range serotype, specific ST subtypes that display a narrower host range than other STs have been identified (3). For example, ST variants DT2 was shown to be associated to pigeons (3). Invasive ST MLST sequence type 313 (ST313) from Sub-Saharan Africa was proposed to may have adapted to human hosts in certain degree (5). Such observations suggest substantial diversity within the population structure of this generalist serotype.

Whole-genome sequencing (WGS) has emerged as a powerful tool for subtyping and phylogenetic analysis (6). Currently WGS-based single nucleotide polymorphisms (wgSNP) typing is the most widely used tool for phylogenetic analysis using WGS data due to its ability to collect phylogenetically informative genetic variations among entire genomes (7, 8). The routine application of WGS in public health surveillance has been creating large volumes of WGS data for major foodborne pathogens. The publicly available WGS data and their associated metadata, such as the GenomeTrakr (NCBI Bioproject#: 183844), make it now possible to do more comprehensive population structure level analysis on prevalent pathogens such as ST.

WGS has been used to study ST populations and their transmission. For example, ST313 is a major cause of invasive salmonellosis in sub-Saharan Africa. Intracontinental spread of ST313 has been investigated using WGS and phylogenetic analysis (5). The investigation estimated that two major lineages of ST313 emerged less than 60 and 40 years ago, which coincided with the current HIV pandemic. It was further proposed that ST313 may have occupied the immunocompromised human population in sub-Saharan Africa as a new niche. In another study, the global transmission of ST DT104 was inferred using WGS data (9). The study indicated that the currently circulating multidrug resistance (MDR) DT104 was likely to be diverged through horizontal transfer of *Salmonella* genomic island 1 around 1970s. Most reported WGS studies of ST populations are focused on clinical strains. A comprehensive investigation of ST isolates from various sources is yet to be performed.

When the population structure of an organism is delineated, comparison among distinct lineages can provide important insights on the biology, evolution and epidemiology of the organism. A notable difference among lineages is the presence or absence of specific genes. For example, Den Bakker, et al. compared the partial gene contents among the genomes of different *Salmonella enterica* serotypes. It was found that the two subpopulations of *S. enterica* had their specific genes. Such genes were involved in mediating *Salmonella* tropism for hosts and tissues as well as the utilization of carbon and nitrogen sources characteristic of certain hosts. Therefore, they were proposed to play a role in the ecology and transmission characteristics of *S. enterica* (10).

Sometimes, more subtle mutations in gene sequence instead of the absence or presence of entire genes can cause phenotypic differences. For example, a SNP that causes a premature stop codon can lead to a loss-of-function truncation of a gene, such as in the case of *inlA* gene to

Listeria monocytogenes, in which a premature stop codon mutation in *inlA* led to production of a truncated InlA which yield virulence-attenuated *Listeria monocytogenes* strains (11). Coding sequences that completely or partially lost its original functions are often called pseudogenes, which contain deficiencies such as frameshifts, premature stop codons, deleterious SNP, etc. The level of pseudogene accumulation has been suggested to be highly related to the process of host adaptation (12-14).Specifically to *Salmonella*, multiple host-restricted or high level host-adapted serotypes such as Typhi and Paratyphi A (human) (13, 15), Dublin (cattle) (16, 17) and Gallinarum (avian) (17, 18) exhibit higher levels of pseudogenes accumulations than serotypes of wide host ranges. Also within a generalist serotype such as Typhimurium, some lineages that show signs of host-adaptation have been reported, such as in the case of DT2 (pigeon).

Microbial source tracking (MST) is a set of genotypic and phenotypic approaches can be utilized to trace fecal pollution and pathogens to their source origin, such as human, cattle, bird, etc. (19). Even though most MST studies are focused on fecal contamination identification of environmental waters (19-21), efforts have been made to utilize MST or quasi-MST methods to trace back the origins of foodborne pathogens in food safety control (22, 23). With the challenges for current MST approaches, such as the requirement of large libraries, low rates of correct classification, etc. (20, 24), WGS has been proposed to address some of them and increase MST performance (25). However, so far, there still haven't been any studies directly applied WGS-based subtyping into MST.

In this chapter, we constructed a wgSNPs based ST phylogeny of isolates from various sources, including clinical strains from the Centers for Disease Control (CDC) and publicly available genomes in the FDA GenomeTrakr database. Distinct population groups or phylogenetic clades associated with food and wild animal sources were identified and

characterized with evolutionary, comparative genomics and phenotypic analyses. In addition, the potential of WGS-based subtyping in MST for foodborne pathogen was demonstrated.

2. Materials and Methods

2.1 Strains and genomes

Genomes of 1,267 ST isolates (between 1930 and 2014) were studied including i) clinical isolates (n=107) from the collection of National *Salmonella* Reference Laboratory at CDC representing diverse pulsed field gel electrophoresis (PFGE) and multiple locus variable number tandem repeat analysis (MLVA) patterns of ST between 1930 and 2014, ii) publicly available genomes from various sources and geographic locations that including 46 US states and 39 countries (n=1,161). Most of these genomes were downloaded from the GenomeTrakr database, including all the ST genomes in the database by September 2015. None duplicate removal was done to maintain an enough sampling intensity, so the over reorientation issue may exist.

According to their sources, the strains can be further categorized into: 1) human (n=152); 2) food (including all types of food except livestock animal meat, n=99); 3) wild birds (n=77); 4) poultry (chicken, turkey and eggs, n=209); 5) bovine (cattle, beef, and dairy product, n=266); 6) swine (pigs and pork, n=247); 7) other land animals (n=174) and 8) others (from any sources other than the aforementioned ones or without specific source information, n=43).

Detailed strain information can be found in Supplementary Table S4.1.

2.2 DNA extraction and sequencing

The 107 clinical isolates were grown in tryptic soy broth (TSB) overnight. Genomic DNA was prepared by using the GenElute Genomic DNA isolation kit (Sigma-Aldrich, St. Louis, MO, USA). The extracted genomic DNA samples were sequenced by an Illumina HiSeq instrument

(San Diego, CA, USA) according to manufacturer's instructions at Translational Genomics Research Institute (Phoenix, AZ).

2.3 WGS-based serotype identification

Any genome of which serotype was not verified to be ST was excluded from further anlysis. All the 1,267 genomes were confirmed to be ST by SeqSero (26).

2.4 *De novo* genome assembly

Trimmomatic (27) was used to trim the raw sequencing reads to eliminate the reads with quality score lower than 20. Then the raw reads were assembled into draft genomes using SPAdes by default settings (28). The quality of each draft assembly was assessed by QUAST (29). Assemblies with an N50 value lower than 100,000 were considered to be poorly assembled and excluded from further analysis.

2.5 Phylogenetic analysis

Strain SL1344 (NCBI accession#: NC_016810) was used as the reference genome for phylogenetic analysis. Any regions involved in repetitive sequences, phages and DNA recombination were detected and removed from the reference genome prior to phylogeny construction. Repeat (longer than 100bp) and phage sequences were detected by repeat-match (30) and Phast (31), respectively, in the genome of SL1344. Recombination regions were inferred from the consensus alignment of all the genomes using Gubbins (32) and ClonalFrameML (33). To generate the alignment, raw reads of each genome were aligned to the reference genome through BWA (34) and high quality SNPs were called using Freebayes (35). High quality SNPs were defined to meet three criteria: 1) at least 5 reads were mapped to the locus of a SNP, 2) at least 75% of mapped reads supported the SNP to be called, and 3) the minimum mapping quality score generated by BWA at the SNP loci is 20. A consensus sequence

for the query genome in comparison with the reference was produced by replacing SNP loci in the reference with the bases called in the query genome and leaving any regions in the reference genome that had less than 5 mapped reads as gaps. The consensus sequences for all the query genomes were combined with the reference genome to form an alignment as the input for Gubbins and ClonalFrameML. Any sequence detected by either tool was considered to be recombination-related. The genomes of G1 group, which will be described in later section, were not included in the recombination analysis due to their large phylogenetic distances to genomes of other groups, which were not optimal for the recombination detection tools.

A maximum likelihood phylogenetic tree was constructed based on the core genome alignment of all the draft and complete ST genomes by Parsnp of the Harvest suite (version 1.2) (36). The modified SL1344 genome (free of inferred repeat, phage and recombination regions) was used as the reference for the alignment. The tree was built by FastTree2 (37) as part of the Parsnp program. To created tree was midpoint rooted by Dendroscope (38).

2.6 Population structure

Major population groups were identified using Bayesian Analysis of Population Structure (BAPS) software package (39) based on the alignment of concatenated SNPs detected by Parsnp. Two levels with a maximum number of 50 populations were set as the initial conditions for BAPS to infer population structure of sampled genomes using a maximum likelihood approach.

To identify any subtrees within the ST ML tree that displayed a strong temporal signal of SNP accumulation, every internal node of the ST ML tree was analyzed by calculating the linear spearman correlation coefficient between isolation years of the corresponding isolates in the subtree (defined by the internal node) and the tip to root (i.e., the internal node) distances of these isolates. Subtrees with a spearman correlation coefficient larger than 0.5 were selected for

further confirmation of the temporal signals. To confirm, the genomes in each selected subtree were first aligned to their nearest complete ST genome on the ML tree of all isolates. The choice of a closely-related reference genome when available maximized the detection of SNPs existing among the members of a subtree but not in a more distant reference genome. High quality SNPs were detected as previously described. Concatenated SNPs were generated from consensus alignment by SNP-site (40). A ML phylogenetic tree was built using concatenated SNPs by Phyml (41). TempEst (previously called Path-O-Gen (42)) was used to calculate the linear coefficients between isolates years and branch lengths of selected subtrees. Any subtree with the coefficient higher than 0.5 was confirmed to have a strong temporal signal and then subjected to modeling-based population dynamics analysis using BEAST 1.8.2 (43) using at least 10,000,000 states by sampling every 5,000. The initial 10% of samples were regarded as burn-in which were removed from further analysis. The maximum clade credibility tree was generated by TreeAnnotator v1.8.2 (http://beast.bio.ed.ac.uk/TreeAnnotator) to estimate the most recent common ancestors. Model performance was assessed through Tracer v1.6 (http://beast.bio.ed.ac.uk/Tracer).

Different combinations of tree models and clock models were tested including combinations 1) Gaussian Markov Random Fields (GMRF) as tree model and relaxed log-normal molecular clock as clock model, 2) GMRF as tree model and strict clock rate as clock model, 3) constant population model as tree model and relaxed log-normal molecular clock as clock model, and 4) constant population model as tree model and strict clock rate as clock model. The combination with the highest Bayesian Factor was selected to do the analysis.

2.7 Statistic test of source-associated clusters

To statistically evaluate the association between a particular source and a specific cluster on the ST ML tree, a previously described test statistic (T) (44) was used. A modification was made to use the number of SNPs between two isolates as a measure of genetic distance. The null hypothesis that no significant clustering of isolates from the same source existed on phylogenetic tree was tested. The test statistic was calculated as follows:

$$T = \sum_{k \in \{\text{all sources}\}} \frac{\sum_{i,j:i \neq j} d_{ij} I(S_i = k) I(S_j = k)}{\sum_{i,j:i \neq j} I(S_i = k) I(S_j = k)}$$

where d_{ij} is the SNP distance between isolates i and j, S_i is the source of isolate i, and $I(S_i = k)$ is an indicator function that returns a value of 1 if isolate i is from source population k. Permutation was performed to calculate the p-value.

2.8 Comparative gene content analysis

Genomes were annotated with Prokka (45). The annotated genomes were analyzed by Roary (46) to compare gene contents among major poulation groups, which was divided into three major categories: 1) pan-genome which refered to the entire gene set of all genomes in the population, 2) core-genome which refered to the genes that were present in most of or all genomes in the population (>=99% genomes), and 3) accessory genome which refered to the genes that were not present in most of or all genomes of the population (< 99% genomes). We defined disparately distributed genes (DDGs) as the genes of which prevalence (percentage of the isolates that had the gene) differed by at least 50% between two major population groups. DDGs were extracted from the output of Roary. Their presence and absence in the genomes of a population group was confirmed by BLAST (47). If a DDG was aligned to a genome and at least 90% of the bases in the DDG were identical to the BLAST hit in the genome (defined as 90% BLAST identity), the DDG was confirmed to be present in the genome. Conversely, the DDG was confirmed to be absent in the genome. Only confirmed DDGs were included for further

analysis. Hierarchical cluster analysis of the DDGs based on their distribution among analyzed genomes was performed using the hclust package of R. Clustered DDGs putatively associated with plasmids, genomic islands or other mobile genomic elements (MGE) were identified through BLAST search against NCBI nucleotide collection (nr/nt). The identified MGEs were confirmed to be present in particular ST genomes by aligning the plasmids or phages to the draft or complete ST genomes by BLAST again with a 90% BLAST identity threshold as previously described.

2.9 Antimicrobial resistance genes (ARG) analysis

The ResFinder (48) database was used to identify acquired ARGs in draft or complete ST genomes. Each allele in the database was aligned to a genome using BLAST. The presence of an ARG allele was determined using default ResFinder setting. Hierarchical cluster analysis of ARGs was performed using the hclust package of R.

2.10 Pseudogene analysis

Potential core-genome pseudogenes were identified by having any of the following mutations: 1) non-synonymous SNPs (NS-SNP) in the start or stop codon; 2) frameshift mutations caused by insertion or deletion (indel); 3) truncations that spanned at least 20% of a coding region; 4) NS-SNPs or non-frameshift indels that were potentially deleterious (explained below); and 5) premature stop codons.

Specifically, indels were identified from raw sequence reads using Scalpel (49) with strain ST SL1344 (NCBI accession#: NC_016810) as the reference genome. Putative indels were first called if 1) at least 5 reads were mapped the locus of an indel, 2) at least 75% of mapped reads supported the indel to be called. Then the high fidelity indels were confirmed by sequence comparison with corresponding sequences in the reference genome using BLAST. The position

of each indel was located in the reference genome. The reference genome sequence between 100 bp upstream and 100 bp downstream of the indel was extracted and compared with the assembled genome of the ST strain in which the indel was identified by BLAST. The BLAST result was examined manually to validate the indel. Deleterious NS-SNPs were identified by Provean (50). A conservative cut-off score of -2.5 was applied. Any NS-SNP with a score lower than -2.5 was considered to be deleterious. To determine truncations to core genes, the predominant allele of each core gene was identified. Any gene with at least 20% of its coding region deleted in comparison with the predominant allele was considered to be truncated. NS-SNPs in start and stop codons and premature stop codons were identified by locating the mutations in the reference genome.

2.11 Metabolic profiling

Overall metabolic potentials of 6 randomly selected strains from 6 major population groups were evaluated with Biolog Phenotype Microarrays (51) Micro Plate 1-4 on different carbon, nitrogen, sulphur, and phosphonate sources. They were sent to Biolog (http://www.biolog.com/) for evaluation. Briefly, the strains were inoculated into each well of the plates. Colorimetric measurements of the wells were taken every 15 min up to 48 hours. Data analysis was initially performed using OmniLog PM Systems. Average signal intensity of each well was calculated based on all time points. Principal component analysis (PCA) was conducted using the Biolog data through prcomp package in R (52).

2.12 Inference of genotypic causes of Biolog phenotypes

For isolates that displayed inability or lower ability to utilize certain substrates in the Biolog Phenotype Microarray analysis as compared with other isolates, the phenotypes were analyzed for their potential genotypic causes. Genes involved in the pathway to utilize a substrate were

identified using Kobas (53) and the KEGG database (54). Any absence or inactivating mutation of these genes, i.e., DDG and pseudogene, were identified as the potential reason for the phenotype. When no evidence was shown to support the phenotype characteristics, the query mutations can be extended to non-synonymous SNP (NS-SNPs).

2.13 Genomes used for evaluating WGS-based source prediction

Two sets of genomes were used to retrospectively evaluate the performance of WGS-based source prediction of ST. First, 17 ST isolates from 8 outbreaks with confirmed origin in swine, poultry and bovine between 1998 and 2016 (Table 4.4), including 5 reported multistate outbreaks, i.e. 2009 pot pie turkey outbreak (55), 2011 ground beef outbreak (56), 2013 ground beef outbreak (57), 2013 live poultry outbreak (58) and 2015 pork outbreak (59). Second, 244 genomes with annotated sources as livestock animals in the FDA GenomeTrakr database network (Supplementary Table S4.2). These genomes were made publicly available after we collected the 1,267 ST genomes to build the ST phylogeny. All the genomes were assembled into draft genomes using SPADES (28).

2.14 WGS-based source prediction of ST

A ML tree was built in the same way described in previous section including both the set of new genomes, i.e. query genomes, and the set of previous 1,267 genomes, i.e. reference genomes. The source prediction for each query genome was made by consulting the source of its nearest reference genome.

2.15 Clades associated with livestock animals

Monophyletic groups of more than 5 reference isolates with a single source accounting for at least 75% of isolates in the groups were identified and scanned from the entire ST ML tree.

2.16 Rarefaction analysis

Rarefaction analysis was performed to compare relative sampling richness of individual sources. Phylogenetic clusters were identified from the ST ML tree. A cluster was defined as a monophyletic group of closely related isolates as measured by the maximum pairwise SNP distance among the members of the group. This heuristic value determined the number of phylogenetic clusters to be identified; the lower the value, the more clusters were found. The maximum pairwise SNP distance of each previously identified livestock animal association clade was calculated and the mean of these distances was used as the heuristic value for phylogenetic cluster identification. Rarefaction curves for particular sources were constructed by plotting the number of phylogenetic clusters as a function of the number of isolates sampled. The vegan package of R (60) was used for drawing rarefaction curves.

3. Results and Discussion

3.1 Phylogeny

A total of 9 separate regions containing 307,970 nucleotide bases were detected to be potential phage or prophage sequences (Supplementary Table S4.3). A total of 154 separate regions containing 269,212 nucleotide bases or 5.6% of a typical ST genome (4.8 Mb) were detected to be potential recombination sequences (Supplementary Table S4.4). Notably, 75.2% of the recombination sequences were associated with phage or prophage. A total of 331 genes had at least 100 bp or 10% of their coding regions identified as recombination sequences (Supplementary Table S4.5**).** A total of 39,562 SNPs were identified from the core genome alignment of 3,978,014 nucleotide bases from the 1267 ST genomes. A ML tree based on these SNPs is shown in Figure 4.1. Based on the two-level BAPS analysis, the sampled ST population were divided into 10 major population groups (G1-G10 in Figure 4.1), among which 8 were monophyletic and 2 were polyphyletic.

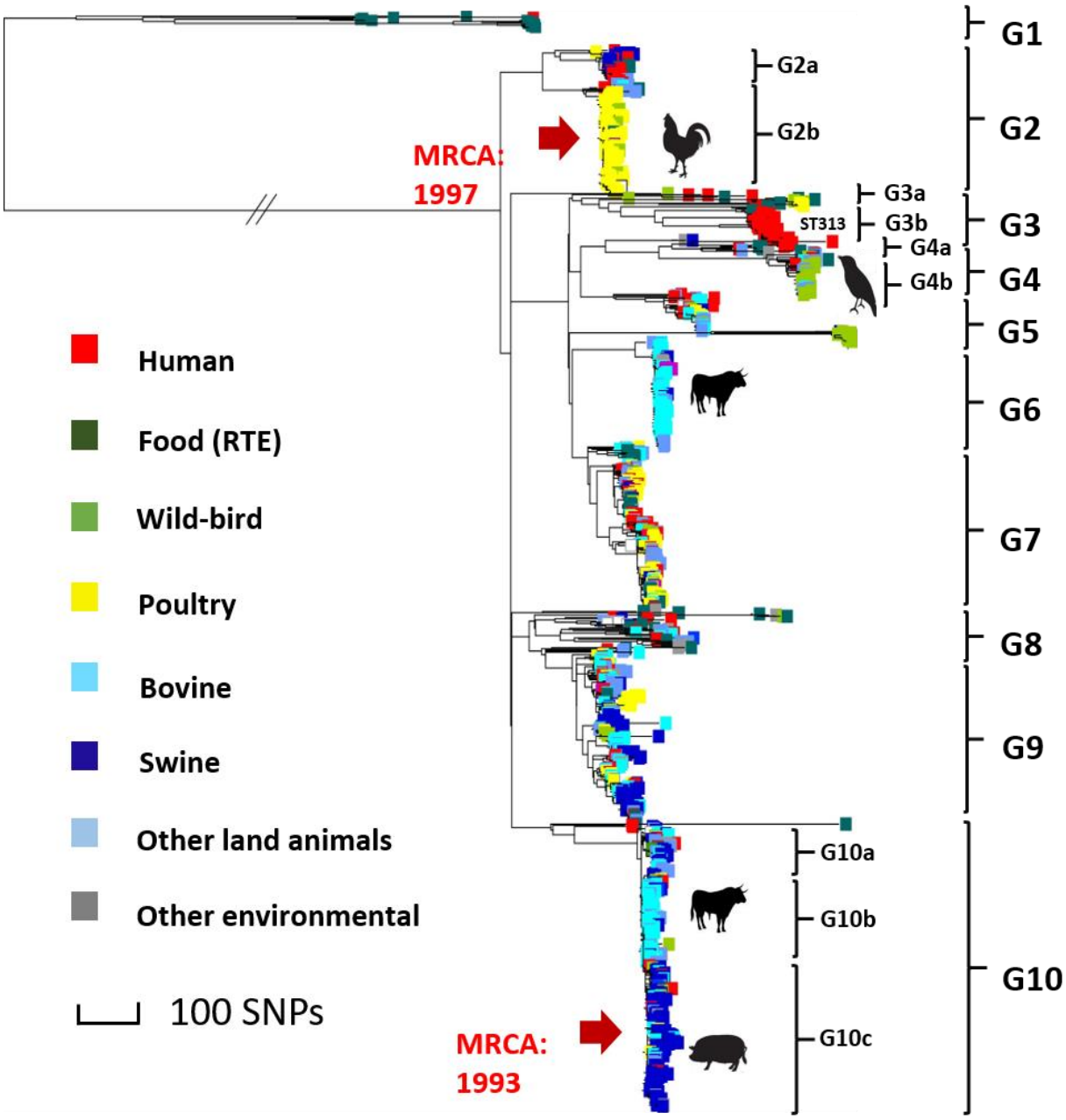

Figure 4.1. The core-genome SNPs based phylogenetic tree.
The estimated ages of most recent common ancestors (MRCA) are shown for two clades. Based on their sources, the strains can be further categorized into Human: clinical isolates (n=152); Food (RTE): isolates from all kinds of food except livestock animal meat and dairy (n=99); Wild-bird: isolates from wild bird (n=77); Poultry: isolates from chicken, turkey and eggs,(n=209); Bovine: isolates from cattle, beef, and dairy product (n=266); Swine: isolates from pigs and pork (n=247); Other land animals: isolates from all land animals except previous categories (n=174); Others: isolates from any sources other than the aforementioned ones or without specific source information (n=43). G1-10 are the population groups identified by BAPS.

3.2 Phylogenetic clustering of isolates from specific sources

In some population groups, clustering of isolates from the same source was observed in multiple cases, including clades overrepresented by isolates from poultry, bovine, swine, sea food (the international group G1), and wild birds. The null hypothesis that no clustering of isolates from any specific source was rejected by applying the statistical model based on comparing mean pair-wise SNP distance between isolates from the same source ($p<0.01$), which suggested significant association existing between particular sources and population groups.

Isolates from livestock animal sources such as bovine, swine and poultry were not evenly distributed on the tree but appeared to cluster in several major population groups compared to other sources (Table 4.1): 88% bovine samples in G6, G9 and G10, 87.9% swine samples in G9 and G10 and 93.3% poultry samples in G2, G7 and G9. Sampled isolates from the three major livestock animal sources also clustered more closely as shown by their smaller than average pairwise SNP distances (Table 4.2).

Clinical and food isolates are widely distributed over the tree and found in almost every major population group with the exception of G6 that did not include a food isolate (Figure 4.1). This observation suggested that ST isolates from various lineages have the potential to be involved in human infections through food vehicles.

Table 4.1. The distributions of genomes from different sources in different population groups

Groups	Sources [a]								
	Bovine	**Porcine**	**Poultry**	**Other-land-animal**	**Clinical**	**Food**	**Wild bird**	**Others**	**Sum**
G1	0	0	0	0	1 (6.0%,0.7%)	15 (94.0%,15.2%)	0	0	16
G2	2 (1.0%,0.8%)	15 (9.0%,6.1%)	111 (66.0%,53.1%)	8 (5.0%,4.6%)	18 (11.0%,11.8%)	6 (4.0%,6.1%)	7 (4.0%,9.1%)	2 (1.0%,4.7%)	169
G3	0	0	4 (8.0%,1.9%)	0	34 (69.0%,22.4%)	9 (18.0%,9.1%)	2 (4.0%,2.6%)	0	49
G4	2 (3.0%,0.8%)	0	0	14 (22.0%,8.0%)	10 (16.0%,6.6%)	6 (10.0%,6.1%)	30 (48.0%,39.0%)	1 (2.0%,2.3%)	63
G5	10 (16.0%,3.8%)	4 (6.0%,1.6%)	5 (8.0%,2.4%)	15 (24.0%,8.6%)	11 (17.0%,7.2%)	1 (2.0%,1.0%)	15 (24.0%,19.5%)	2 (3.0%,4.7%)	63
G6	94 (76.0%,35.3%)	5 (4.0%,2.0%)	0	12 (10.0%,6.9%)	1 (1.0%,0.7%)	0	0	12 (10.0%,27.9%)	124
G7	14 (7.0%,5.3%)	4 (2.0%,1.6%)	58 (30.0%,27.8%)	32 (16.0%,18.4%)	35 (18.0%,23.0%)	38 (19.0%,38.4%)	7 (4.0%,9.1%)	7 (4.0%,16.3%)	195
G8	4 (8.0%,1.5%)	2 (4.0%,0.8%)	0	12 (23.0%,6.9%)	11 (21.0%,7.2%)	12 (23.0%,12.1%)	2 (4.0%,2.6%)	9 (17.0%,20.9%)	52
G9	39 (19.0%,14.7%)	54 (27.0%,21.9%)	26 (13.0%,12.4%)	41 (20.0%,23.6%)	20 (10.0%,13.2%)	8 (4.0%,8.1%)	8 (4.0%,10.4%)	7 (3.0%,16.3%)	203
G10	101 (30.0%,38.0%)	163 (49.0%,66.0%)	5 (2.0%,2.4%)	40 (12.0%,23.0%)	11 (3.0%,7.2%)	4 (1.0%,4.0%)	6 (2.0%,7.8%)	3 (1.0%,7.0%)	333
Sum	266	247	209	174	152	99	77	43	1267

[a] The two percentages in the brackets: former one is the percentage in population groups (calculated by row) and the latter one is the percentage in specific sources (calculated by column). For example, G2-Poultry cell is "111 (66.0%, 53.1%)" which means 111 isolates in G2 are from Poultry source, and they are 66.0% of all G2 isolates and 53.1% of all Poultry source isolates

Table 4.2. Mean pairwise SNP distance among isolates sampled from individual sources

Sources	Mean	STD	Min	Median	Max
Bovine	**378.11**	230.81	0	483	881
Porcine	**297.94**	235.78	2	443	955
Poultry	**336.64**	224.50	2	458	850
Clinical	649.12*	1101.16	0	527	10333 *
Food (RTE)	2938.31*	4132.55	0	561	10268 *
Wild birds	632.79	321.33	2	751	1178
Others	478.97	194.43	0	504	1094
All sources	721.4304	1481.994	0	516	10372

* The abnormal high mean-SNPs and max-SNPs were caused by G1 group isolates which are far from other STs. If removed G1 isolates, the mean-SNPs for Clinical and Food (RTE) are 524 and 487, and the max-SNPs for Clinical and Food (RTE) are 936 and 1252.

Overall, six major population groups appeared to be overrepresented by isolates from a specific source. In such cases, isolates from a single source accounted for over 40% of all isolates in the population group. These examples included seafood and plant in G1, poultry in G2, human in G3, wild bird in G4, bovine in G6, and bovine and swine together in G10. By comparison, no individual sources stood out in the other four major population groups (G5, G7, G8 and G9).

G1 (n=16) is phylogenetically distant from other population groups that appears to diverge from other population group early in the history. Isolates in G1 were separated by 9,790-10,047 SNPs from the reference genome from strain SL1344 (NCBI accession#: NC_016810), similar to typical SNP distances among different serotypes. Most of the isolates were sampled from food category in Asia, including seafood (n=10 out of 16, 62.5%) and plant-sourced food (n=5 out of 16, 31.25%. The small size of this group (about 1.26% of whole population), and geographic association to Asia and large distance to other major population groups sampled in the US suggested this is not a typical ST lineage circulating in the US.

G2, G6 and G10 are three population groups over-represented by isolates sampled from livestock animals. G2 (n=169) can be further divided into two clades: the smaller one (G2a, n=44) mainly made up of strains from swine (n=15, 34.1% of G2a) and human (n=17, 38.6% of G2a), and the bigger one (G2b, n=125) associated with poultry (n=110, 88% of G2b) (Figure 4.1). The high percentage of clinical strains in G2a, compared with that of 12% across the entire tree, suggested that this clade might have been frequently involved in human infections. G6 (n=124) appeared to be associated with bovine (n=94, 76% of G6) and G10 (n=333, the largest group) appeared to be associated with both bovine and swine (bovine n=101, 30% of G10; swine n=163, 49% of G10). G10 can be further divided into three clades: G10a, G10b and G10c (Figure 4.1). G10a (n=58) consisted of isolates from bovine (n=12, 20.1%), swine (n=20, 34.5%) and other land animals (n=16, 27.6%). G10b (n=108) was over-represented by bovine isolates (n=78, 72.2%) and G10c (n=167) by swine isolates (n=133, 79.6%).

G3 (n=49) was a geographically diverse group with isolates from multiple regions in the world, including sub-Saharan Africa, US, China, Mexico, India, and European countries. The major component of this group was ST313 strains from humans (n=32, 65.3%), a novel MLST sequence type circulating in sub-Saharan Africa region and often causing invasive systemic disease (5). Two clades were identified in G3, including G3a that had non-ST313 isolates (n=17) and G3b that had ST313 isolates (n=32). Isolates that had a most recent common ancestor with the ST313 clade were of the MLST sequence type ST302, which had been reported in Mexico and Zimbabwe (61).

G4 (n=63) was also a population group with isolates from multiple countries including US, UK, Canada, Egypt and Sweden. Isolates from wild birds were overrepresented in G4 (n=30, 48% of G4). Other sources that were found in this group included humans, plants and other land

animals, i.e. the animals except swine, bovine and poultry. Two major clades were identified in G4 (Figure 4.1). G4a included all non-wild-bird isolates in this group (n=13) and G4b was the major wild bird clade in the US (n=50) that included wild bird isolates (n=30) from states of New York, Maryland, Minnesota, Idaho and Washington.

G5, G7, G8 and G9 were regarded as diverse-source-groups due to the diverse origins of these groups' isolates and the lack of a single dominant source in these groups. G5 (n=63) and G8 (n=52) are two polyphyletic groups, while G7 (n=195) and G9 (n=203) are monophyletic.

Among these groups, G7 and G8 featured high percentages of clinical and food isolates. While food and clinical sources only accounted for 7.8% and 12% of all the sampled isolates in this study (n=1,267), in G7 and G8 18% and 21% of isolates were from clinical cases, and 19% and 23% were from food samples. The seemingly overrepresentation of clinical and food isolates in G7 and G8 suggested that these two population groups were commonly circulating in the US and often associated with food contamination and human infection. The presence of isolates from all the other sources defined in this study further suggest a wide distribution of the two population groups in various environments.

3.3 Clades of temporal signal in SNP accumulation

Three clade of moderate temporal signals in SNP accumulation as indicated by a correlation coefficient >= 0.5 were found in G2b, G3b and G10c. Such temporal signals allowed us to date the emergence of each clade under a Bayesian temporal framework. The first clade in G2b was associated with poultry. Bayesian population analysis showed that the most recent common ancestor (MRCA) of this clade lived around year 1997 with a 7.3 * 10^-7 substitutions per site per year. The second one was a ST313 clade in G3b, which was reported to be divided into two clades: one diverged around 1960 with 1.9 * 10^-7 substitutions per site per year and the other

diverged around 1972 with 3.9 * 10^-7 substitutions per site per year (5). The third clade included the majority of swine isolates in G10c (n=128, 76.7% of G10c). It was inferred that the MRCA of this cluster emerged around 1993 with an average mutation rate at 5.1 * 10^-7 substitutions per site per year. Unlike other major *Salmonella* serotypes such as Enteritidis and Heidelberg, where significant temporal signals can be observed from the entire phylogenetic trees of isolates from various sources (62, 63), in ST only certain clades of the same source displayed such signals. The livestock animal ST clades all had a young MRCA less than 24 years old, implying recent establishment of the clades in a particular niche.

3.4 Comparative genomics analysis

3.4.1 Gene contents

A total of 24,817 genes (pan-genome) were identified from all the analyzed ST genomes, among which a total of 7,955 genes were present in at least 1% genomes. The core genome and accessory genome of this population contained 3,795 genes and 21,022 genes respectively. A total of 1,104 disparately distributed genes (DDGs) were found from core genome between population groups (Supplementary Table S4.6). Hierarchical clustering of accessory genes based on their presence and absence among analyzed genomes identified genes specific to particular population groups or clades (Figure 4.2 cluster A-D):

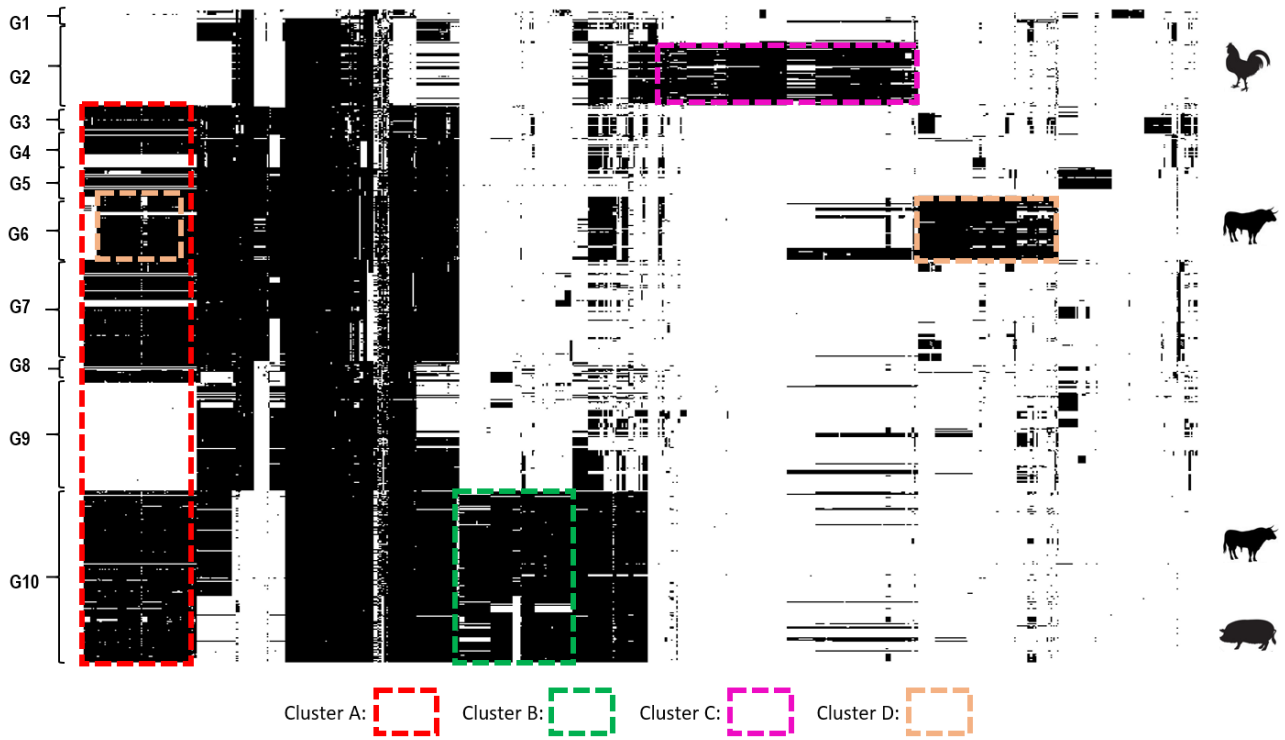

Figure 4.2. Distribution of Disparately Distributed Genes (DDG) among ST genomes.

Each black point stands for the presence of one gene. The order of genomes is the same as that of the ST phylogenetic tree.

Cluster A is the virulence plasmid pSLT in ST that carries the *spv* genes. These genes play a role in ST infection of spleen and liver of mouse by increasing the rate of bacterial replication within host cells (64, 65).

Cluster B included *Salmonella* genomic island 1 (SGI1) (66), a virulence gene *pipB2* that can alter host cell physiology and promote bacterial survival in host tissues (67), and some prophage sequences. SGI1 contains multidrug resistance regions (68), which have been reported to favor G10's circulation among cattle and pigs under the selection pressure of agricultural antibiotics in animal farming (68).

Cluster C was specific to poultry-associated G2b and contained a mega plasmid (220 kb) with multiple antibiotic resistance genes. The plasmid had been isolated from chicken breast in the US

(69). Cluster D was specific to bovine-associated G6. The cluster included a plasmid that has recently been reported from cattle in Japan (70). This plasmid (pYT2) was derived from the virulence plasmid pSLT by acquiring an at least 21 kb region that harbors a resistance island and is flanked by IS1294 elements. The two plasmids' putative association with poultry or bovine source and their multidrug resistance genes suggest that they may contribute to the adaption of certain ST clades to certain farm environments.

Besides mobile genetic elements such as plasmids, we also examined the distribution of antibiotic resistance genes (ARG) among all the population groups (Figure 4.3). Clustering of multiple types of ARG based on their presence or absence among ST genomes was apparent in livestock animal associated groups such as G2a (swine and human), G2b (poultry), G4b (human), G6 (bovine) and G10 (bovine and swine). Moreover, most of these groups appeared to have a characteristic combination of ARGs (Figure 4.3). In comparison, population groups of diverse sources (G5, G7, G8 and G9) and associated with wild birds (G4) seemed to carry fewer ARG (Figure 4.3). These observations suggest that agricultural use of antibiotics may have an impact on the ARG composition of ST isolates from food animal production and such impact may vary among different livestock environments.

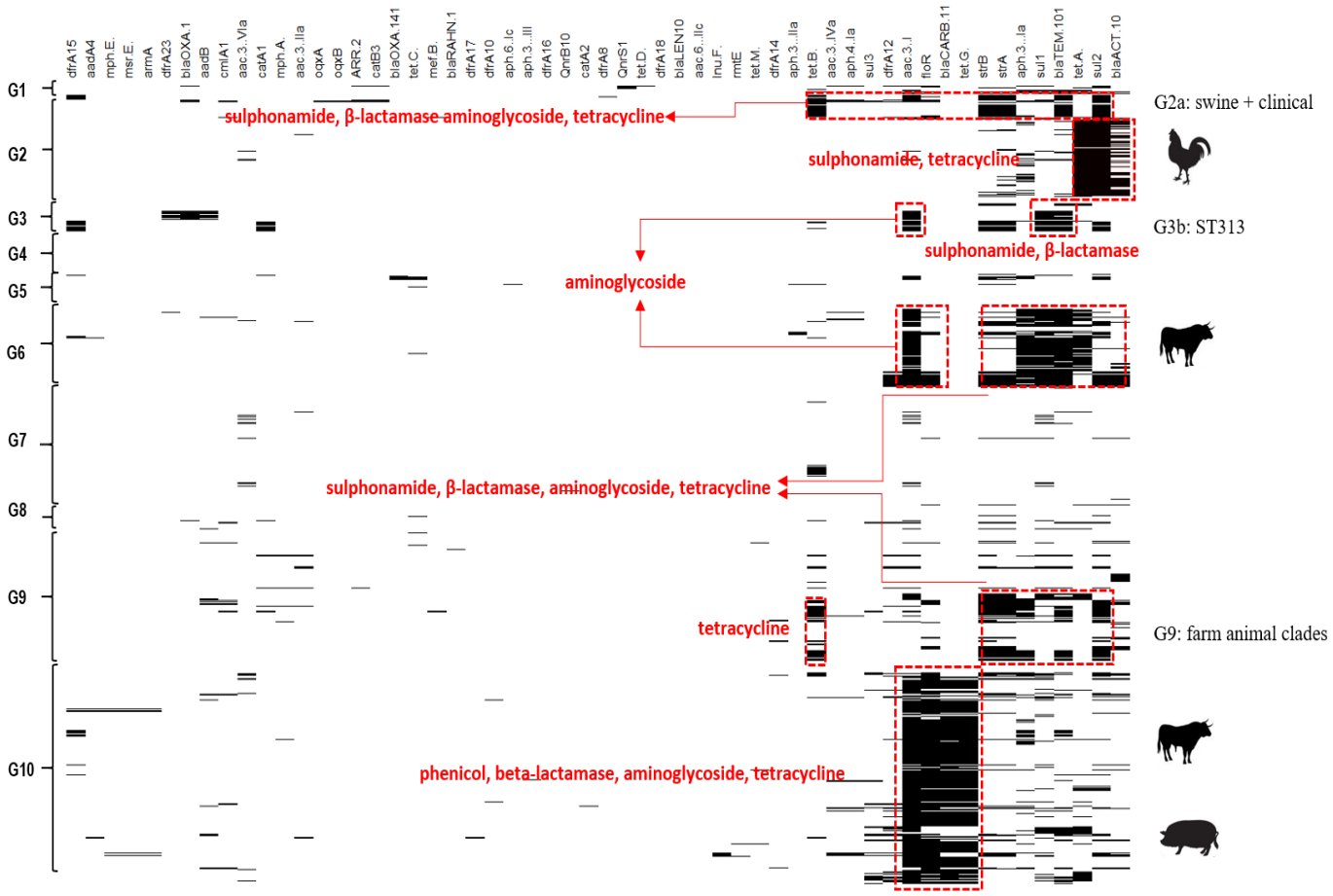

Figure 4.3. Distribution of acquired antibiotic resistance genes among ST genomes. Each black dash stands for the presence of one gene. The vertical order of genomes are same as phylogenetic tree. The labels at the top are the names of AMRs. The clustering of ARG were not in the whole G9 group but in its livestock animal clades.

3.4.2 Pseudogene analysis

A total of 844 potential core-genome pseudogenes were identified (Supplementary Table S4.7). Similar to that of DDGs in accessory genome, the distribution of potential pseudogenes reflected the population structure of sampled ST genomes, with most of the population groups displaying a characteristic pattern of pseudogene content (Figure 4.4).

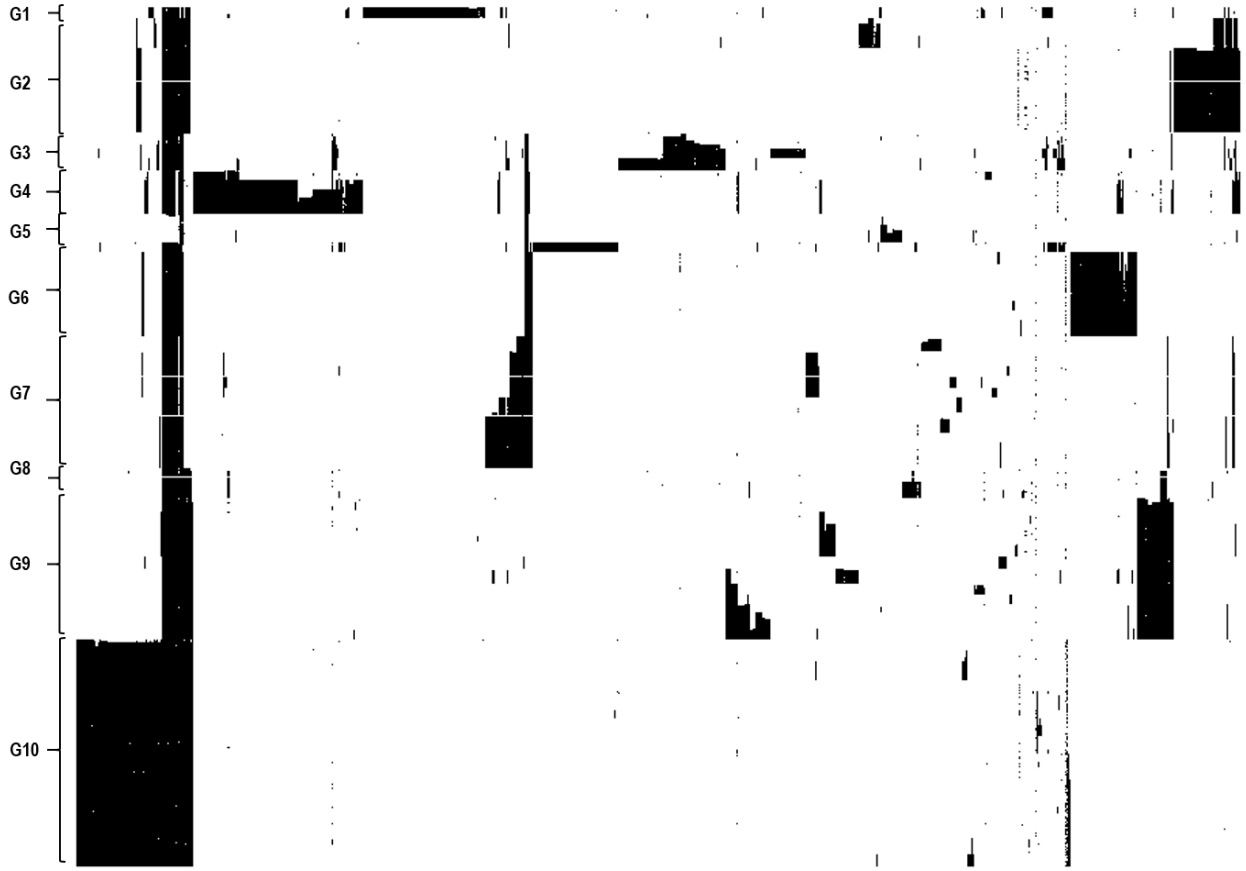

Figure 4.4. Distribution of potential pseudogenes among ST populations.
Each black point stands for the presence of a potential pseudogene. The order of genomes is the same as that of the ST phylogenetic tree.

The numbers of potential pseudogenes in ST genomes of different population groups were compared (Figure 4.5). Potential pseudogenes were most abundant in the genomes of G4b (wild birds), G1 (Asian seafood and plants) and G3b (ST313), followed by those of G10 (bovine and swine), G2b (poultry) and G6 (bovine) that were associated with livestock animals. The genomes of diverse-source groups had fewest potential pseudogenes; the median number of pseudogenes in each of these group was lower than that across all sampled genomes and groups (Figure 4.5). The accumulation of pseudogenes or genome degradation has been proposed as an indicator –of bacterial host-adaption. For example, an evaluated number of pseudogenes were detected for Typhi and Paratyphi A to human (13) and Gallinarum to poultry (71). Within the serotype of

Typhimurium, multiple clones displaying different host ranges have been identified (3). Epidemiological evidence in Germany suggested that pigeon and wild bird-associated ST definitive phage types (DT2) were highly host-adapted (3). The relatively high degree of genome degradation in the wild bird group observed in this study provided further evidence that some ST lineages have adapted to the avian hosts. ST313 is another ST subtype that has been reported to have accumulated pseudogenes as a sign of host adaptation (3). Our finding confirmed the large number of pseudogenes in this lineage as compared with other population groups of ST.

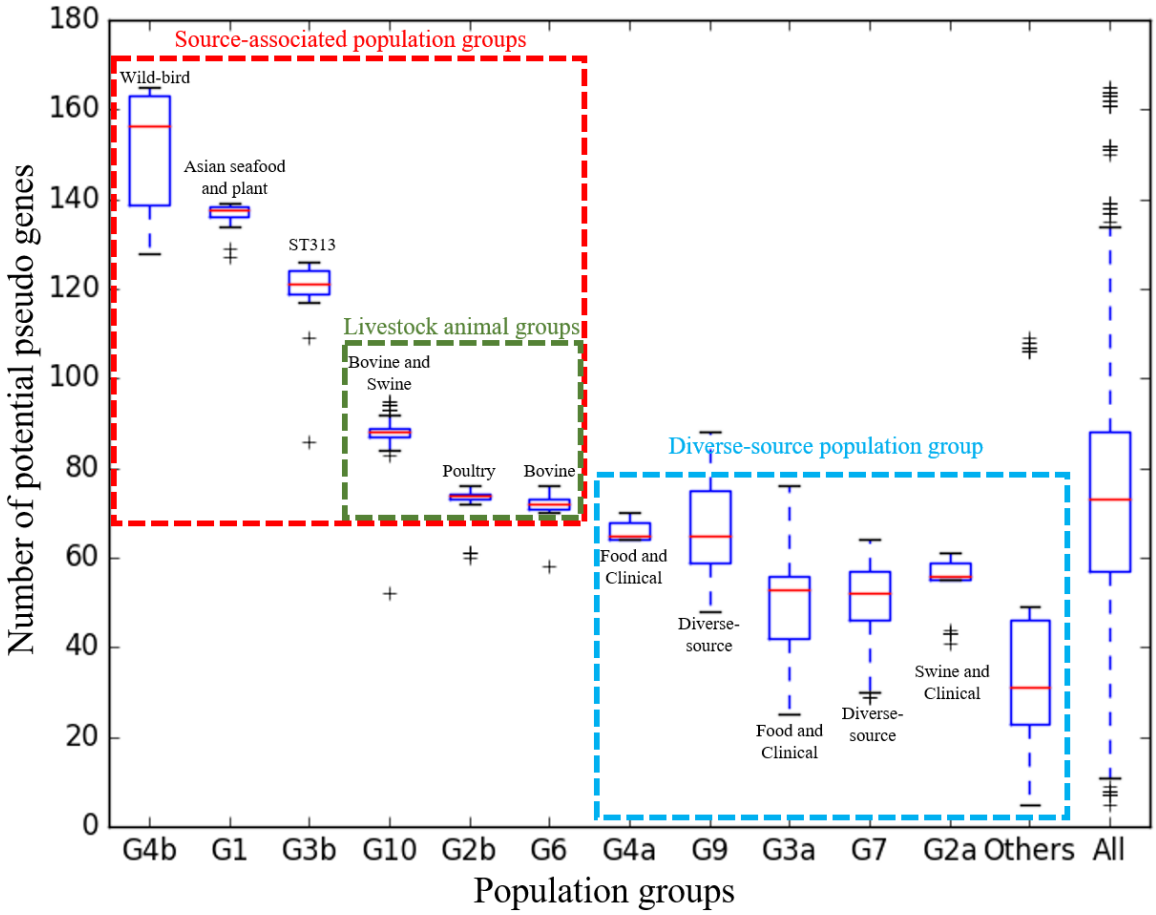

Figure 4.5. Box-plots about potential pseudogenes frequencies on different population groups.

The box of each group stands for upper and lower quartiles. The height of each box represents the distance between upper and lower quartiles, i.e. interquartile range (IQR). The red dash in the middle of each box means median. The lower end of each whisker means the lower quantile minus 1.5 IQR. The upper end of each whisker means the upper quantile plus 1.5 IQR. "+" means the outliers not between two whiskers.

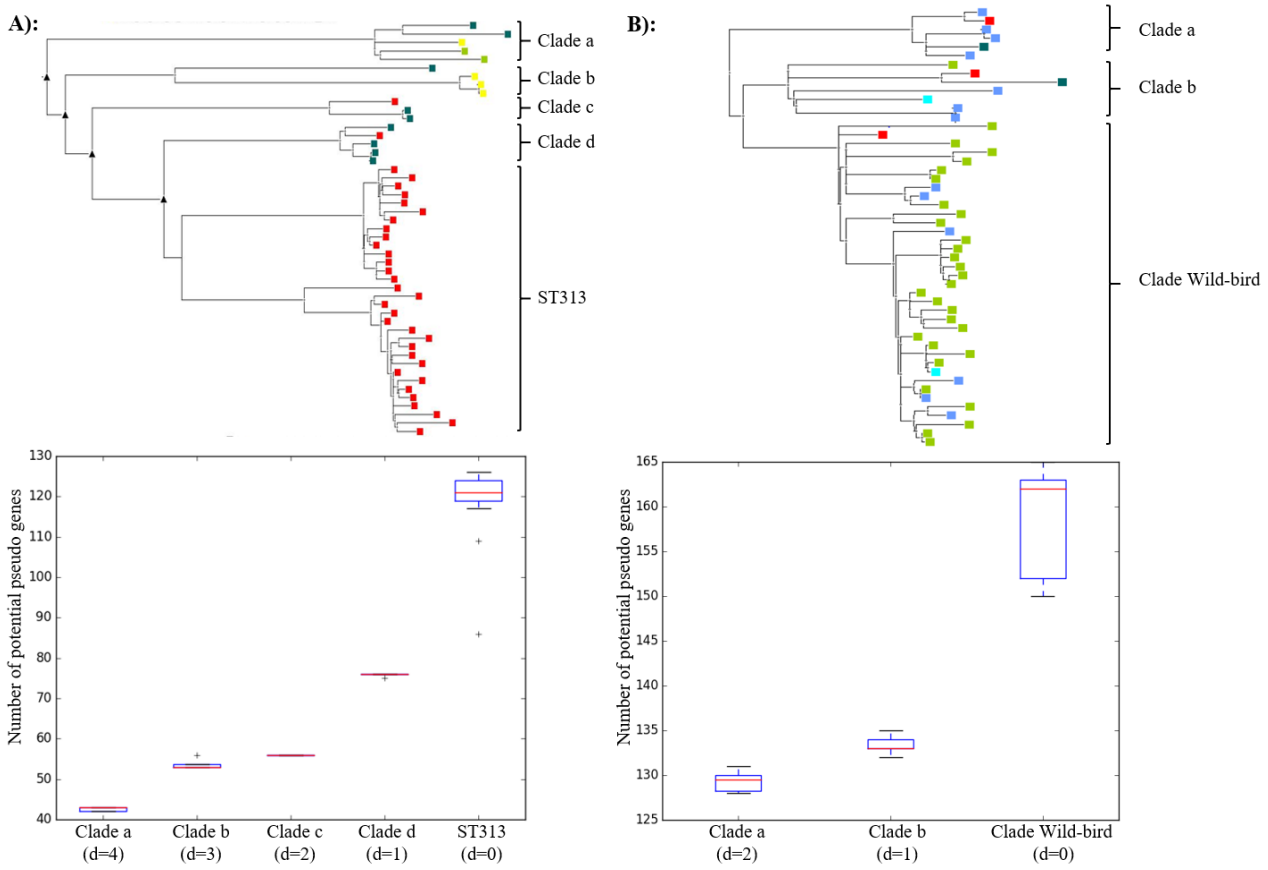

Figure 4.6. Accumulation of pseudogenes as theST313 (G3b) and Wild bird (G4b) clades diverged.

d is the number of internal ancestral nodes (divergence events) between tested clade and target clade (ST313 or Wild bird clade).

3.5 Comparative analysis of metabolic profiles

3.5.1 Phenotype characteristics of major population groups

Metabolic potentials of representative isolates from different sources were characterized by evaluating the isolates' abilities to utilize carbon, nitrogen, phosphorus and sulfur sources using Biolog Phenotype Microarray Plates 1-4 (Figure 4.7). These isolates represented all the monophyletic US population groups (G2, G4, G6, G7, G9 and G10) identified in the study.

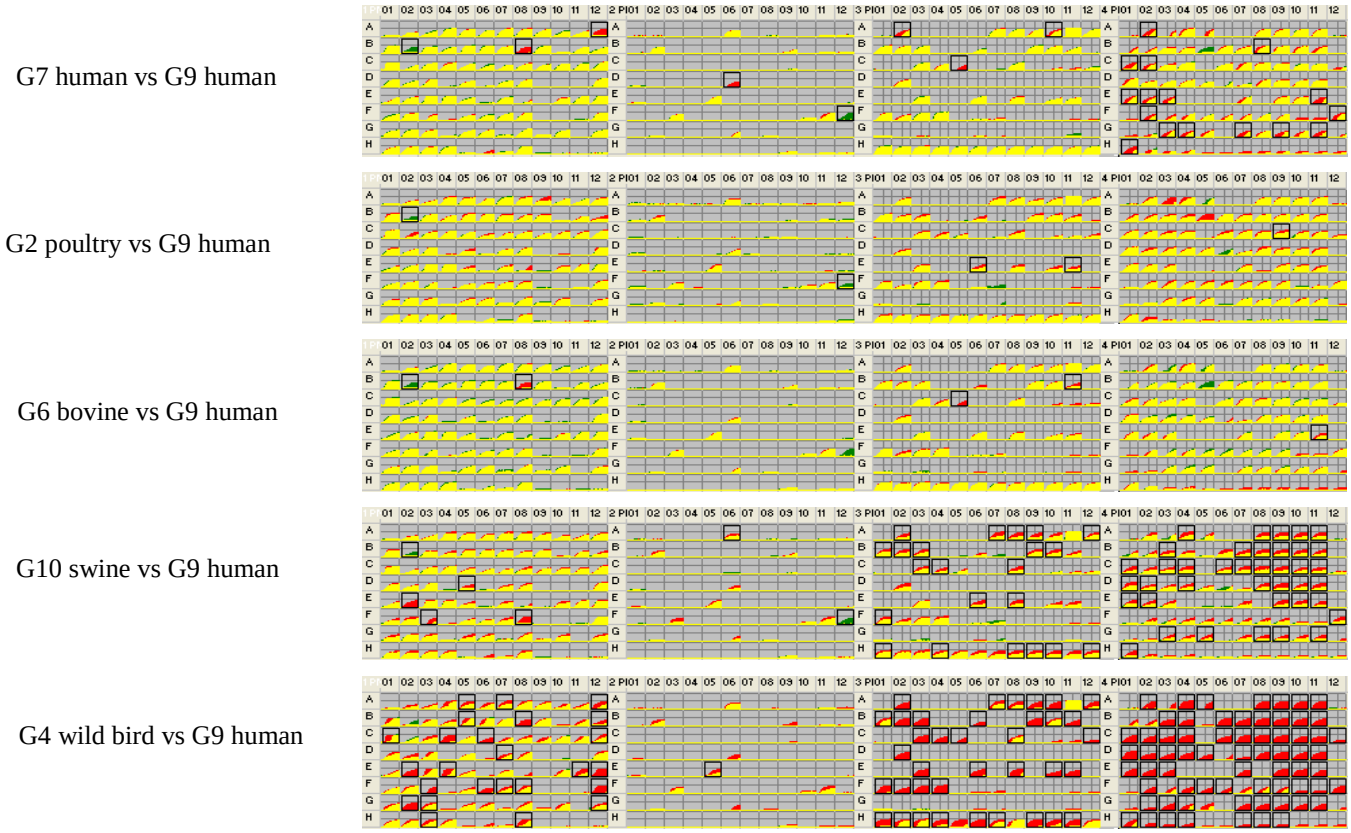

Figure 4.7. Phenotype Microarray results.
G9 (Human) as reference (red) and the others as targets (green), the overlap part of reference and target was shown yellow. From left to right, the four plates are: PM1 (Carbon), PM2 (Carbon), PM3 (Nitrogen) and PM4 (A-E: Phosphorus; F-H: Sulfur).

Different metabolic phenotypes were observed among some of the representative isolates, especially with respect to the utilization of nitrogen, phosphorus and sulfur. Principal component analysis based on all the tested substrates revealed that representative isolates from poultry (G2), bovine (G6), human (G7 and G9) had similar metabolic potentials, while the swine (G10) and wild bird (G4) isolates each displayed an overall metabolic profile distinct from any other isolate (Figure 4.8). In comparison with other isolates, the swine isolate and the wild bird isolate displayed moderately and severely reduced ability in utilizing substrates that provided carbon, nitrogen, phosphorous and sulfur source, respectively. In the case of the wild bird isolate, its

apparent restricted metabolic potential is in particular agreement with the high level of genome degradation identified in this population group (Figure 4.5), further suggesting a narrow host range.

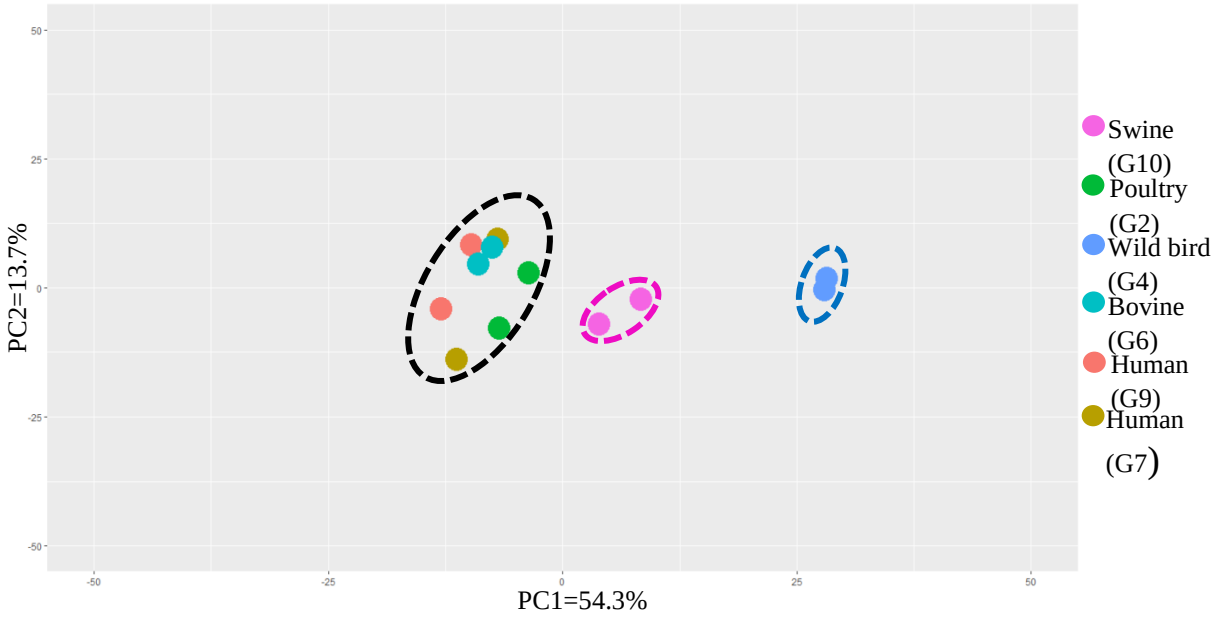

Figure 4.8. Principal component analysis (PCA) of metabolic profiles of representative isolates from each monophyletic population group.
Results from 2 replicate assays were shown.

Similarly, the swine isolate from G10 had more potential pseudogenes than other livestock animal groups (Figure 4.5). While showing a lesser degree of genome degradation and narrowed substrate utilization range compared with the wild bird isolate, the swine isolate did provide evidence for a certain level of host adaptation of G10.

3.5.2 Potential genotype evidence that may explain observed metabolic phenotypes

Co-analysis of the potential pseudogenes, DDGs and non-synonymous SNPs (NS-SNPs) and the Biolog Phenotype Microarray data yielded potential gene targets to which mutations may affect metabolic pathways involved in utilizing certain substrates (Table 4.3). The NS-SNPs

listed in Table 4.3 were not considered to be deleterious as opposed to the NS-SNPs categorized as a type of potential pseudogenes (See Methods for details).

The affected metabolic pathways in the wild bird isolate (G4b) had more pseudogenes (6 out of 13) than swine isolate (3 out of 12), whereas non-deleterious NS-SNPs accounted for the majority of mutations (9 out of 12) in the swine isolate (G10). This difference may partially explain why the bird isolate was often incapable while the swine isolate was typically ineffective in utilizing corresponding substrates.

Table 4.3. Potential genotypes affecting phenotype microarray cells [a]

Isolates	Gene	Type	Nutrient Sources	Affected PM Cells	Reference Genome Locus [b]
G4 (wild-bird)	pstA	Psedo_gene	P	No growth from PM4-A1 to E12	SL1344_3822
G4 (wild-bird)	cysN	Psedo_gene	S	No growth of PM4-F2	SL1344_2913
G4 (wild-bird)	yecS	NS-SNP	S	Low growth of PM4-F7,F8,G3,G4,H1	SL1344_1566
G4 (wild-bird)	pepA	NS-SNP	S	Low growth of PM4-F9,G5	SL1344_4407
G4 (wild-bird)	narY	NS-SNP	N	Low or no growth of PM3 cells	SL1344_1509
G4 (wild-bird)	narW	NS-SNP	N	Low or no growth of PM3 cells	SL1344_1510
G4 (wild-bird)	glnD	NS-SNP	N	Low or no growth of PM3 cells	SL1344_0215
G4 (wild-bird)	patD (ydcW)	Psedo_gene	N	No growth of PM3-C12	SL1344_1528
G4 (wild-bird)	livH	Psedo_gene	N; C	No growth of PM3-B3,B11; PM2A-G6	SL1344_3529
G4 (wild-bird)	dtpB (yhiP)	Psedo_gene	N	Low or no growth of PM3-H1 to H12	SL1344_3557
G4 (wild-bird)	rbsA	NS-SNP	C	No growth of PM1-B8,C4	SL1344_3849
G4 (wild-bird)	iolG2	Psedo_gene	C	No growth of PM1-F3	SL1344_4363
G4 (wild-bird)	tktA	NS-SNP	C	No growth of PM1-B8,C4,E12,F12	SL1344_2310
G10 (Swine)	frwA	NS-SNP	P	Low growth from PM4-A1 to E12	SL1344_4060
G10 (Swine)	ydhU (phsC)	NS-SNP	S	Low growth PM4-F3,F4	SL1344_2039
G10 (Swine)	moaA	NS-SNP	S	Low growth from PM4-F2 to H12	SL1344_0778
G10 (Swine)	thiI	NS-SNP	S	Low growth from PM4-F2 to H13	SL1344_0419
G10 (Swine)	gcvP	Pseudo_gene	N	Low growth PM3-B2,B10,B11	SL1344_3029
G10 (Swine)	dsdA	Pseudo_gene	N	Low growth PM3-B2,B10,B11	SL1344_3770
G10 (Swine)	sdaA	NS-SNP	N	Low growth PM3-B2,B10,B11	SL1344_1755
G10 (Swine)	narV	NS-SNP	N	Low or no growth of PM3 cells	SL1344_1511
G10 (Swine)	gltL	NS-SNP	N;C	Low or no growth of PM3 cells; PM1-A7,B12	SL1344_0651
G10 (Swine)	gltI	NS-SNP	N;C	Low or no growth of PM3 cells; PM1-A7,B12	SL1344_0654
G10 (Swine)	livM	NS-SNP	N; C	Low growth of PM3-B3,B11; PM2A-G6	SL1344_3529
G10 (Swine)	gcl	Pseudo_gene	C	Low growth of PM1-E2	SL1344_0510

[a] Details about the metabolic pathways are in Supplementary Table S4.8
[b] The reference genome is Strain SL1344 (NCBI accession#: NC_016810)

3.6 WGS-based MST of ST

Two sets of genomes were used to evaluate how sampled ST isolates (n=1,267) and their population structure could support MST of ST. To make a source prediction, each query genome

was incorporated into the ST ML tree and the source of the closest reference genome (i.e., one of the 1,267 genomes) to the query was used to predict the source of the query genome.

A total of 8 ST outbreaks with confirmed origin in livestock animals and related meat and dairy products occurred between 1998 and 2015 in the United States. Among them, isolate source was correctly predicted for 5 outbreaks (Table 4.3). For these correct predictions, all the reference genomes were isolated at least 2 years before or after the outbreak, suggesting that they were epidemiologically unrelated to the outbreaks. The temporal separations between the isolation of reference isolates and the occurrence of the outbreaks indicated the persistence of the source-associated ST populations that eventually caused human infections.

Two Isolates from a beef outbreak were closely related to a clade that had both a poultry and a bovine reference isolates, with the poultry isolate being slightly closer to the beef outbreak isolates (Figure 4.9C). Similarly, two isolates from a live poultry outbreak had a recent common progenitor with a clade consisting of a bovine and a poultry reference isolates, with the bovine reference isolate being slightly closer to the poultry outbreak isolates (Figure 4.9B). It is possible that the outbreak isolates involved in the two cases were able to circulate among both cattle and chicken hosts or production environments. The small number of reference isolates (n=2) that were closely related to the query isolates from each outbreak also contributed to the ambiguity in source prediction.

The isolate from the 2015 pork outbreak was within a clade of human clinical isolates without further source information (Figure 4.9A). Therefore, source prediction was not made for this isolate.

Among the outbreaks whose sources were correctly predicted, one was caused by source-associated lineages, i.e. 2011 ground beef outbreak to G6. The rest of the 4 outbreaks were

attributed to the diverse-source population groups G3, G2a and G8. Successful prediction of isolate source for these outbreaks suggested that within a diverse-source population group, individual clades linked to particular sources can be identified and used for source prediction.

A total of 244 isolates of annotated livestock animal sources (poultry, swine and bovine) had been deposited to the FDA GenomeTrakr database before April 1st 2016 and after we initially collected the 1,267 ST genomes. Correct source prediction was made for 180 isolates (73.8%), of which 160 were from source-associated population groups and 20 from diverse-source population groups (Supplementary Table S4.2).

Table 4.4. The metadata and prediction results of CDC confirmed outbreak-isolates

Isolates [#]	CDC confirmed outbreaks		Major lineage	Nearest reference genomes			Prediction
	Outbreak	Year		Isolates	Source	Year	
SRR5201365	2013 ground beef multistate outbreak	2013	G3	Lab#68	poultry	2005	Correct
SRR5201363	2013 ground beef multistate outbreak	2013	G3	Lab#68	poultry	2005	Correct
SRR5201332	2009 pot pie turkey multistate outbreak	2009	G3	Lab#96	poultry	2005	Correct
SRR5201423	2009 pot pie turkey multistate outbreak	2009	G3	Lab#96	poultry	2005	Correct
SRR5201390	2009 pot pie turkey multistate outbreak	2009	G3	Lab#96	poultry	2005	Correct
SRR5201382	2009 pot pie turkey multistate outbreak	2009	G3	Lab#96	poultry	2005	Correct
SRR5201385	2013 live poultry multistate outbreak	2013	G3	Lab#38	bovine	2005	Incorrect
SRR5201393	2013 live poultry multistate outbreak	2013	G3	Lab#38	bovine	2005	Incorrect
SRR5201371	2011 ground beef multistate outbreak	2011	G6	Lab#291	bovine	2007	Incorrect
SRR5201361	2011 ground beef multistate outbreak	2011	G6	Lab#291	bovine	2007	Incorrect
SRR2194006	2015 pork multistate outbreak	2015	G2a	NA*	NA	NA	Incorrect
Lab# 1055	Cattle-contact-verified	2010	G8	Lab#336	bovine	2008	Correct
Lab# 1114	Cattle-contact-verified	2010	G2a	Lab#1017	bovine	2010	Correct
Lab# 1034	Cattle-contact-verified	2010	G2a	Lab#1017	bovine	2010	Correct
Lab# 1035	Cattle-contact-verified	2010	G2a	Lab#1017	bovine	2010	Correct
Lab# 1104	Farm-rawmilk outbreak	1998	G8	Lab#36	bovine	2004	Correct
Lab# 1027	chicken verified	2009	G8	Lab#125	poultry	2003	Correct

[#] Accession# or Lab#

[*] There is no source reference genome in the clade

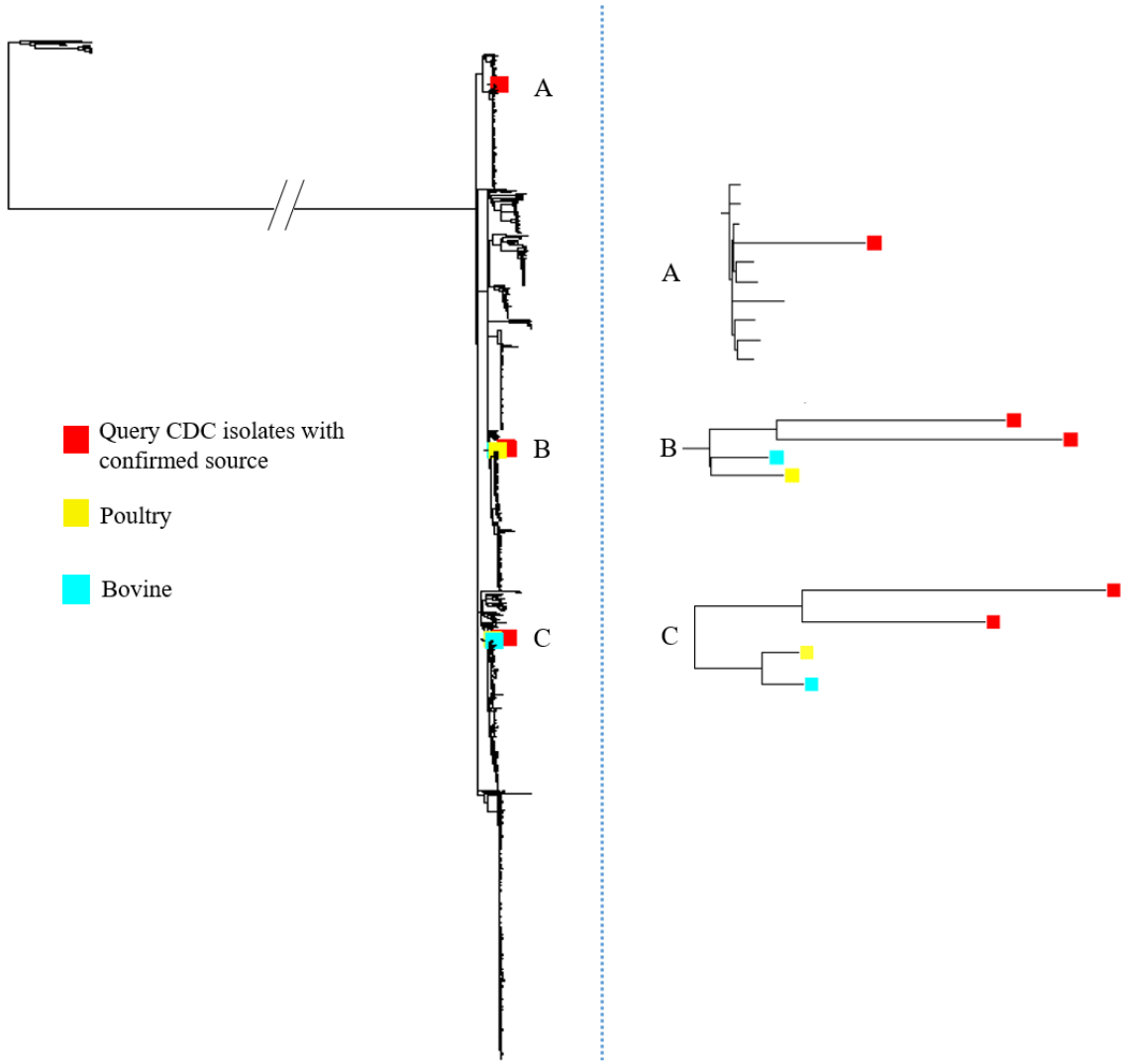

Figure 4.9. Phylogenetic clades containing outbreak isolates without correct source prediction.

All the leaf nodes in Figure 4.9A without a square node shape are clinical genomes without confirmed sources, which was not included in MST analysis.

3.7 Differential sample intensity of ST reference genomes

A major factor that determines the performance of WGS-based source prediction of ST is the amount and diversity of reference ST genomes. As more genomes become available and more source-associated ST clades are represented, it is more likely to find closely-related reference genomes for robust source prediction of query genomes.

Rarefaction analysis was used to examine the number of phylogenetic clades represented by a certain number of sampled isolates. A proper definition of phylogenetic clades for this analysis should reflect the observed phylogenetic diversity within individual source-associated clusters. A total of 21 livestock animal associated clusters were identified from the 1,267 sampled isolates. For each of these clusters, we calculated the maximum SNP distance between any pair of genomes as a quantitative measure of phylogenetic diversity within the cluster (See Methods for details). The distribution of these maximum in-cluster pairwise SNP distances had a mean value of 73 SNPs. This mean value was used as an empirical cutoff to algorithmically identify distinct clades across the entire tree such that each identified clade had a maximum in-clade pairwise SNP distance no larger than 73 SNPs. A total of 267 clades were delineated from the ST ML tree of 1,267 isolates.

As shown in Figure 4.10, differential sampling intensities were detected among different categories of isolates by source. For livestock and wild bird categories, the curves displayed lower slopes which indicated relatively higher sampling intensities, whereas the clinical and food categories showed steeper slopes suggesting that the phylogenetic diversity of clinical and food isolates was less sampled than that of the isolates from wild birds and livestock animals. This observation may be explained by a multitude of origins and routes leading to food contamination and human infection, which made the isolates from these two sources inherently diverse.

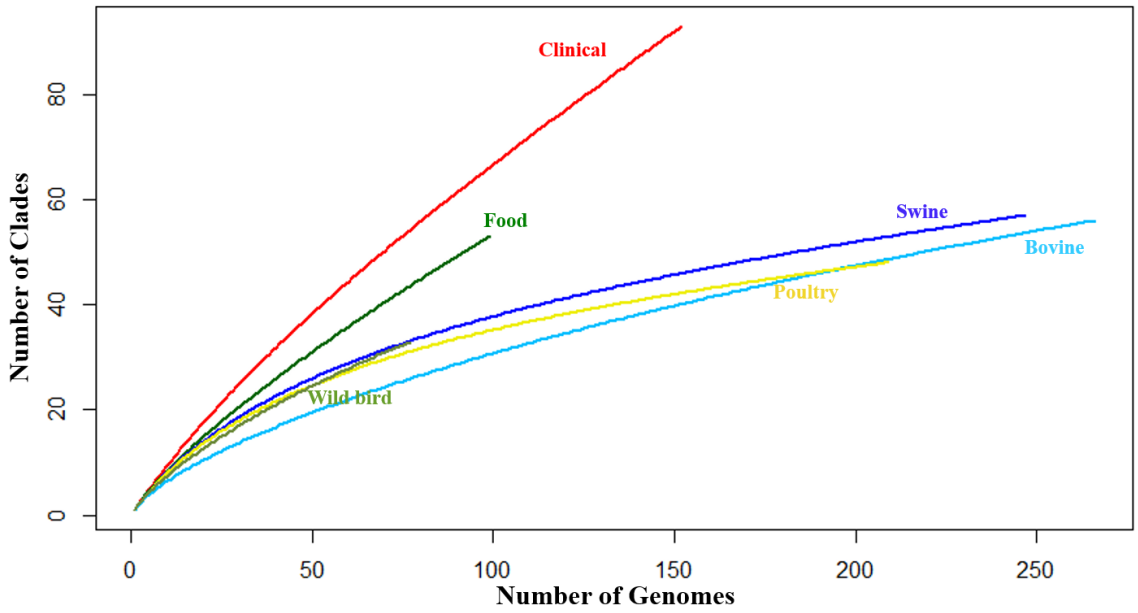

Figure 4.10. The rarefaction curves for reference genomes of different source categories. All isolates from same category are displayed in same color, i.e. clinical isolates in red, food isolates in dark green, swine isolates in dark blue, bovine isolates in light blue, poultry isolates in yellow, and wild bird isolates in olive green.

4. Conclusion

ST is generally regarded as a broad-host-range serotype found from multiple sources [4]. However, some specific ST subpopulations were reported to show a narrow host range, for example DT2 to pigeon (3). In this study, we showed that both diverse-source population groups that circulate among different sources and source-associated groups that are overrepresented by isolates from a specific source exist in ST populations in the US. The identification of these groups was accompanied by a host of evolutionary, genetic and phenotypic evidence. For example, a wild bird lineage featured a relatively high level of genome degradation and a unique metabolic profile characterized by inability to catabolize multiple metabolites, both indicative of host-adaptation.

One potential application of this study is WGS-based source prediction or MST for ST. We performed an initial evaluation of MST using genomes from livestock sources, including those from actual outbreaks and publicly available in the GenomeTrakr database. While a simple method of only consulting the source of the closest genome to a query genome was used for prediction, our results suggest that this approach is promising. Optimization of the prediction algorithm and continuous expanding of the ST reference genome database are expected to further improve the performance of source prediction.

References

1. CDC.gov. (n.d.) National Enteric Disease Surveillance: *Salmonella* Annual Report, 2013. Retrieved January 20, 2017, from https://www.cdc.gov/nationalsurveillance/pdfs/*Salmonella*-annual-report-2013-508c.pdf
2. Uzzau S, Brown DJ, Wallis T, Rubino S, Leori G, Bernard S, Casadesus J, Platt DJ, Olsen JE. 2000. Host adapted serotypes of *Salmonella enterica*. Epidemiol Infect 125:229-55.
3. Rabsch W, Andrews HL, Kingsley RA, Prager R, Tschape H, Adams LG, Baumler AJ. 2002. *Salmonella enterica* serotype Typhimurium and its host-adapted variants. Infect Immun 70:2249-55.
4. Kingsley RA, Baumler AJ. 2000. Host adaptation and the emergence of infectious disease: the *Salmonella* paradigm. Mol Microbiol 36:1006-14.
5. Okoro CK, Kingsley RA, Connor TR, Harris SR, Parry CM, Al-Mashhadani MN, Kariuki S, Msefula CL, Gordon MA, de Pinna E, Wain J, Heyderman RS, Obaro S, Alonso PL, Mandomando I, MacLennan CA, Tapia MD, Levine MM, Tennant SM, Parkhill J, Dougan G. 2012. Intracontinental spread of human invasive *Salmonella* Typhimurium pathovariants in sub-Saharan Africa. Nat Genet 44:1215-21.
6. Deng X, den Bakker HC, Hendriksen RS. 2016. Genomic Epidemiology: Whole-Genome-Sequencing–Powered Surveillance and Outbreak Investigation of Foodborne Bacterial Pathogens. Annu Rev Food Sci Technol 7:353-374.
7. Aslam ML, Bastiaansen JW, Elferink MG, Megens HJ, Crooijmans RP, Blomberg LA, Fleischer RC, Van Tassell CP, Sonstegard TS, Schroeder SG. 2012. Whole genome SNP

discovery and analysis of genetic diversity in Turkey (*Meleagris gallopavo*). BMC genomics 13:391.

8. Govindaraj M, Vetriventhan M, Srinivasan M. 2015. Importance of genetic diversity assessment in crop plants and its recent advances: an overview of its analytical perspectives. Genet Res Int 2015.

9. Leekitcharoenphon P, Hendriksen RS, Le Hello S, Weill FX, Baggesen DL, Jun S-R, Ussery DW, Lund O, Crook DW, Wilson DJ, Aarestrup FM. 2016. Global Genomic Epidemiology of *Salmonella enterica* Serovar Typhimurium DT104. Appl Environ Microbiol 82:2516-2526.

10. den Bakker HC, Switt AIM, Govoni G, Cummings CA, Ranieri ML, Degoricija L, Hoelzer K, Rodriguez-Rivera LD, Brown S, Bolchacova E. 2011. Genome sequencing reveals diversification of virulence factor content and possible host adaptation in distinct subpopulations of *Salmonella enterica*. BMC genomics 12:425.

11. Nightingale KK, Ivy RA, Ho AJ, Fortes ED, Njaa BL, Peters RM, Wiedmann M. 2008. inlA premature stop codons are common among Listeria monocytogenes isolates from foods and yield virulence-attenuated strains that confer protection against fully virulent strains. Appl Environ Microbiol 74:6570-83.

12. Foley SL, Johnson TJ, Ricke SC, Nayak R, Danzeisen J. 2013. *Salmonella* pathogenicity and host adaptation in chicken-associated serovars. Microbiol Mol Biol Rev 77:582-607.

13. Holt KE, Thomson NR, Wain J, Langridge GC, Hasan R, Bhutta ZA, Quail MA, Norbertczak H, Walker D, Simmonds M, White B, Bason N, Mungall K, Dougan G, Parkhill J. 2009. Pseudogene accumulation in the evolutionary histories of *Salmonella enterica* serovars Paratyphi A and Typhi. BMC Genomics 10:36.

14. Rosinski-Chupin I, Sauvage E, Mairey B, Mangenot S, Ma L, Da Cunha V, Rusniok C, Bouchier C, Barbe V, Glaser P. 2013. Reductive evolution in Streptococcus agalactiae and the emergence of a host adapted lineage. BMC Genomics 14:252.

15. McClelland M, Sanderson KE, Clifton SW, Latreille P, Porwollik S, Sabo A, Meyer R, Bieri T, Ozersky P, McLellan M, Harkins CR, Wang C, Nguyen C, Berghoff A, Elliott G, Kohlberg S, Strong C, Du F, Carter J, Kremizki C, Layman D, Leonard S, Sun H, Fulton L, Nash W, Miner T, Minx P, Delehaunty K, Fronick C, Magrini V, Nhan M, Warren W, Florea L, Spieth J, Wilson RK. 2004. Comparison of genome degradation in Paratyphi A and Typhi, human-restricted serovars of *Salmonella enterica* that cause typhoid. Nat Genet 36:1268-74.

16. Betancor L, Yim L, Martinez A, Fookes M, Sasias S, Schelotto F, Thomson N, Maskell D, Chabalgoity JA. 2012. Genomic Comparison of the Closely Related *Salmonella enterica* Serovars Enteritidis and Dublin. Open Microbiol J 6:5-13.

17. Langridge GC, Fookes M, Connor TR, Feltwell T, Feasey N, Parsons BN, Seth-Smith HM, Barquist L, Stedman A, Humphrey T. 2015. Patterns of genome evolution that have accompanied host adaptation in *Salmonella*. Proc Natl Acad Sci U S A 112:863-868.

18. Feng Y, Johnston RN, Liu GR, Liu SL. 2013. Genomic comparison between *Salmonella* Gallinarum and Pullorum: differential pseudogene formation under common host restriction. PLoS One 8:e59427.

19. Graves AK. 2011. Food safety and implications for microbial source tracking, p 585-607, Microbial Source Tracking: Methods, Applications, and Case Studies. Springer.

20. Fu LL, Li JR. 2014. Microbial source tracking: a tool for identifying sources of microbial contamination in the food chain. Crit Rev Food Sci Nutr 54:699-707.

21. Simpson JM, Santo Domingo JW, Reasoner DJ. 2002. Microbial source tracking: state of the science. Environ Sci Technol 36:5279-88.

22. Nayak R, Stewart-King T. 2008. Molecular epidemiological analysis and microbial source tracking of *Salmonella enterica* serovars in a preharvest turkey production environment. Foodborne Pathog Dis 5:115-26.

23. Cooley M, Carychao D, Crawford-Miksza L, Jay MT, Myers C, Rose C, Keys C, Farrar J, Mandrell RE. 2007. Incidence and tracking of Escherichia coli O157:H7 in a major produce production region in California. PLoS One 2:e1159.

24. Scott TM, Rose JB, Jenkins TM, Farrah SR, Lukasik J. 2002. Microbial source tracking: current methodology and future directions. Appl Environ Microbiol 68:5796-5803.

25. Sadowsky MJ, Call DR, Santo Domingo JW. 2007. The future of microbial source tracking studies, p 235-277, Microbial Source Tracking. American Society of Microbiology.

26. Zhang S, Yin Y, Jones MB, Zhang Z, Deatherage Kaiser BL, Dinsmore BA, Fitzgerald C, Fields PI, Deng X. 2015. *Salmonella* serotype determination utilizing high-throughput genome sequencing data. J Clin Microbiol 53:1685-92.

27. Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114-20.

28. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol 19:455-77.

29. Gurevich A, Saveliev V, Vyahhi N, Tesler G. 2013. QUAST: quality assessment tool for genome assemblies. Bioinformatics 29:1072-5.

30. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL. 2004. Versatile and open software for comparing large genomes. Genome Biol 5:R12.

31. Zhou Y, Liang Y, Lynch KH, Dennis JJ, Wishart DS. 2011. PHAST: a fast phage search tool. Nucleic Acids Res 39:W347-52.

32. Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA, Bentley SD, Parkhill J, Harris SR. 2015. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. Nucleic Acids Res 43:e15.

33. Didelot X, Wilson DJ. 2015. ClonalFrameML: efficient inference of recombination in whole bacterial genomes. PLoS Comput Biol 11:e1004041.

34. Li H, Durbin R. 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 26:589-95.

35. Garrison E, Marth G. 2012. Haplotype-based variant detection from short-read sequencing. ArXiv e-prints. 1207.

36. Treangen TJ, Ondov BD, Koren S, Phillippy AM. 2014. The Harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. Genome Biol 15:524.

37. Price MN, Dehal PS, Arkin AP. 2010. FastTree 2--approximately maximum-likelihood trees for large alignments. PLoS One 5:e9490.

38. Huson DH, Richter DC, Rausch C, Dezulian T, Franz M, Rupp R. 2007. Dendroscope: An interactive viewer for large phylogenetic trees. BMC Bioinformatics 8:460.

39. Corander J, Waldmann P, Sillanpää MJ. 2003. Bayesian analysis of genetic differentiation between populations. Genetics 163:367-374.

40. Page AJ, Taylor B, Delaney AJ, Soares J, Seemann T, Keane JA, Harris SR. 2016. SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. bioRxiv doi:10.1101/038190.

41. Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol 52:696-704.

42. Rambaut A, Lam TT, Carvalho LM, Pybus OG. 2016. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). Virus Evolution 2:vew007.

43. Drummond AJ, Suchard MA, Xie D, Rambaut A. 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol Biol Evol 29:1969-1973.

44. Nightingale KK, Lyles K, Ayodele M, Jalan P, Nielsen R, Wiedmann M. 2006. Novel method to identify source-associated phylogenetic clustering shows that Listeria monocytogenes includes niche-adapted clonal groups with distinct ecological preferences. J Clin Microbiol 44:3742-51.

45. Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics 30:2068-9.

46. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MT, Fookes M, Falush D, Keane JA, Parkhill J. 2015. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics 31:3691-3.

47. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. BMC Bioinformatics 10:421.

48. Kleinheinz KA, Joensen KG, Larsen MV. 2014. Applying the ResFinder and VirulenceFinder web-services for easy identification of acquired antibiotic resistance and E. coli virulence genes in bacteriophage and prophage nucleotide sequences. Bacteriophage 4:e27943.

49. Narzisi G, O'rawe JA, Iossifov I, Fang H, Lee Y-h, Wang Z, Wu Y, Lyon GJ, Wigler M, Schatz MC. 2014. Accurate de novo and transmitted indel detection in exome-capture data using microassembly. Nature methods 11:1033-1036.

50. Choi Y, Chan AP. 2015. PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. Bioinformatics:btv195.

51. Shea A, Wolcott M, Daefler S, Rozak DA. 2012. Biolog phenotype microarrays. Microbial Systems Biology: Methods and Protocols:331-373. Springer.

52. Ihaka R, Gentleman R. 1996. R: a language for data analysis and graphics. J Comput Graph Stat 5:299-314.

53. Xie C, Mao X, Huang J, Ding Y, Wu J, Dong S, Kong L, Gao G, Li CY, Wei L. 2011. KOBAS 2.0: a web server for annotation and identification of enriched pathways and diseases. Nucleic Acids Res 39:W316-22.

54. Kanehisa M, Goto S. 2000. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28:27-30.

55. CDC.gov. (2007, October 29) Multistate Outbreak of *Salmonella* I 4,[5],12:i- Infections Linked to Pot Pies (FINAL UPDATE). Retrieved January 30, 2017, from https://www.cdc.gov/*Salmonella*/2007/pot-pie-10-29-2007.html.

56. CDC.gov. (2012, February 1) Multistate Outbreak of *Salmonella* Typhimurium Infections Linked to Ground Beef (Final Update). Retrieved January 30, 2017, from https://www.cdc.gov/*Salmonella*/2011/ground-beef-2-1-2012.html.

57. CDC.gov. (2013, March 15) Multistate Outbreak of *Salmonella* Typhimurium Infections Linked to Ground Beef (Final Update). Retrieved January 30, 2017, from https://www.cdc.gov/*Salmonella*/typhimurium-01-13/index.html.

58. CDC.gov. (2013, November 1) Multistate Outbreak of Human *Salmonella* Typhimurium Infections Linked to Live Poultry in Backyard Flocks (Final Update). Retrieved January 30, 2017, from https://www.cdc.gov/*Salmonella*/typhimurium-live-poultry-04-13/index.html.

59. CDC.gov. (2015, December 2) Multistate Outbreak of Multidrug-Resistant *Salmonella* I 4,[5],12:i:- and *Salmonella* Infantis Infections Linked to Pork (Final Update). Retrieved January 30, 2017, from https://www.cdc.gov/*Salmonella*/pork-08-15/index.html.

60. Oksanen J, Blanchet FG, Kindt R, Legendre P, Minchin PR, O'Hara RB, Simpson GL, Solymos P, Stevens MHH, Wagner H. 2015. Vegan: community ecology package. R package vegan, vers. 2.2-1.

61. Vinuesa P, Puente JL, Calva E, Zaidi MB, Silva C. 2016. Complete genome sequence of *Salmonella enterica* serovar Typhimurium strain SO3 (sequence type 302) isolated from a baby with meningitis in Mexico. Genome Announc 4:e00285-16.

62. Deng X, Desai PT, den Bakker HC, Mikoleit M, Tolar B, Trees E, Hendriksen RS, Frye JG, Porwollik S, Weimer BC, Wiedmann M, Weinstock GM, Fields PI, McClelland M. 2014. Genomic epidemiology of *Salmonella enterica* serotype Enteritidis based on population structure of prevalent lineages. Emerg Infect Dis 20:1481-9.

63. Hoffmann M, Zhao S, Pettengill J, Luo Y, Monday SR, Abbott J, Ayers SL, Cinar HN, Muruvanda T, Li C, Allard MW, Whichard J, Meng J, Brown EW, McDermott PF. 2014. Comparative Genomic Analysis and Virulence Differences in Closely Related *Salmonella enterica* Serotype Heidelberg Isolates from Humans, Retail Meats, and Animals. Genome Biol Evol 6:1046-1068.

64. Gulig PA, Doyle TJ. 1993. The *Salmonella* typhimurium virulence plasmid increases the growth rate of *Salmonella*e in mice. Infect Immun 61:504-511.

65. Rotger R, Casadesús J. 2010. The virulence plasmids of *Salmonella*. Int Microbiol 2:177-184.

66. Doublet B, Boyd D, Mulvey MR, Cloeckaert A. 2005. The *Salmonella* genomic island 1 is an integrative mobilizable element. Mol Microbiol 55:1911-1924.

67. Baisón-Olmo F, Cardenal-Muñoz E, Ramos-Morales F. 2012. PipB2 is a substrate of the *Salmonella* pathogenicity island 1-encoded type III secretion system. Biochem Biophys Res Commun 423:240-246.

68. Beutlich J, Jahn S, Malorny B, Hauser E, Huhn S, Schroeter A, Rodicio MR, Appel B, Threlfall J, Mevius D, Helmuth R, Guerra B, Med-Vet-Net WPPG. 2011. Antimicrobial resistance and virulence determinants in European *Salmonella* genomic island 1-positive *Salmonella enterica* isolates from different origins. Appl Environ Microbiol 77:5655-64.

69. Hoffmann M, Muruvanda T, Allard MW, Korlach J, Roberts RJ, Timme R, Payne J, McDermott PF, Evans P, Meng J. 2013. Complete genome sequence of a multidrug-resistant *Salmonella enterica* serovar Typhimurium var. 5− strain isolated from chicken breast. Genome Announc 1:e01068-13.

70. Tamamura Y, Tanaka K, Akiba M, Kanno T, Hatama S, Ishihara R, Uchida I. 2013. Complete nucleotide sequences of virulence-resistance plasmids carried by emerging multidrug-resistant *Salmonella enterica* Serovar Typhimurium isolated from cattle in Hokkaido, Japan. PloS one 8:e77644.

71. Langridge GC, Fookes M, Connor TR, Feltwell T, Feasey N, Parsons BN, Seth-Smith HMB, Barquist L, Stedman A, Humphrey T, Wigley P, Peters SE, Maskell DJ, Corander J, Chabalgoity JA, Barrow P, Parkhill J, Dougan G, Thomson NR. 2015. Patterns of genome evolution that have accompanied host adaptation in *Salmonella*. Proc Natl Acad Sci U S A 112:863-868.

CHAPTER 5

SUMMARY

In this thesis, we demonstrated the potential of whole genome sequencing (WGS) in *Salmonella* serotyping and microbial source tracking (MST). Firstly, we developed a bioinformatics tool (SeqSero, http://www.denglab.info/SeqSero) which allows accurate, fast and comprehensive predictions on *Salmonella* serotypes based on both WGS raw sequencing reads and draft genome assemblies. SeqSero provides an alternative for current *Salmonella* serotyping approaches which can help to bridge the gap between the well-established utility of *Salmonella* serotyping and the WGS-based pathogen subtyping and characterization. Secondly, we revealed the population structure based on whole genome sequencing SNPs (wgSNPs) of *Salmonella enterica* serovar Typhimurium (ST), one broad-host-range pathogen. Association between certain population groups/clades and some specific sources were observed, based on which the feasibility of WGS-based MST were demonstrated by evaluating one recent ST outbreak dataset and one GenomeTrakr livestock isolate dataset.

There is no denying that limitations existed during our research. For example, for SeqSero, serotype predictions were made by the determination of antigenic profiles of O and H antigens. However, for some serotypes which shares antigenic profiles, such as *Salmonella enterica* serovar Diguel and Telelkebir both in "13:d:e,n,z15", ambiguity can be generated. In future, we will explore the genomes of those serotypes to find additional genomic markers to solve the issue. For ST population structure and MST, one unavoidable problem for any population structure analyses is the sampling level. The failure of MST for three recent ST outbreaks

(Figure 4.9) and the observation of low sampling intensities for clinical and food isolates (Figure 4.10) both pointed that more isolates are required to represent the diversity of true ST population. A good sign is that, with the routine sequencing of microbial pathogen genomes, a boosted number of WGS genomes can be expected in publicly available databases. For example, by the date January 1st 2017, the number of ST genomes in GenomeTrakr has been more than 2,100, which almost doubled compared with the date we initiated the study, September 2015. In future, more available WGS data will be integrated into the ST population analysis and MST study.

APPENDIX

Due to the large size of supplementary tables, all of them were uploaded to public available websites which can be achieved from either www.denglab.info/static/All-Supplementary-tables_Shaokang_Zhang.xlsx or https://github.com/ShaokangZhang/graduate_supplementary_tables