The Impact of Genomics Implementation on a National Enterics Reference and Surveillance Laboratory

By

Christopher Yachison

A Thesis Submitted to the Faculty of Graduate Studies of

The University of Manitoba

in Partial Fulfilment of the Requirements of the Degree of

MASTER OF SCIENCE

Department of Medical Microbiology and Infectious Diseases

University of Manitoba

Winnipeg

Abstract

Foodborne disease surveillance in Canada is a multifaceted endeavor and human clinical disease is captured by two major laboratory-based surveillance systems, which monitor bacterial enteric pathogens, including *Salmonella*. For years, these programs have relied on multiple methodologies for the identification and subtyping of these bacterial pathogens. Recently, public health officials have been looking towards whole genome sequencing (WGS) to provide a large dataset from which all the relevant and predictable information about an isolate can be mined, thereby replacing traditional laboratory-based surveillance methodologies. Rigorous validation and careful implementation plans are required before WGS-based analyses can replace traditional subtyping tests. The objectives of this research were to assess the ability of WGS-based analysis methods to replace traditional subtyping tests used in laboratory-based surveillance systems and to develop the considerations that can guide future implementation of WGS technology. *Salmonella* serotyping remains the gold-standard tool for the classification of *Salmonella* isolates and the *Salmonella in silico* Typing Resource (SISTR) was validated using a set of 492 clinical isolates to uncover gaps in prediction algorithms and improve tool development. Following further refinement, traditional serotyping was compared to SISTR along with another *in silico* serotyping tool, SeqSero, and 7-gene multilocus sequence typing (MLST) for serovar prediction, which can be adapted for *in silico* analysis. Successful results were obtained for 94.8%, 88.2%, and 88.3% of the 813 clinical and laboratory isolates tested using SISTR, SeqSero, and 7-gene MLST, respectively. While the three methods vary in their algorithms, all would be suitable for maintaining the historical records, surveillance systems, and communication structures currently in place. Additionally, two WGS-based phylogenetic analysis methods, the Single Nucleotide Variant Phylogenomics (SNVPhyl) pipeline and whole

genome MLST, were used to retrospectively assess a well characterized multi-serovar outbreak of *Salmonella* from 2014. While both platforms differ in their approaches for analyzing the WGS data, both methods reached the same conclusions and provided an increased resolution to the outbreak investigation. Additional non-outbreak clusters of disease were identified in the WGS analysis, and other PFGE-based clusters identified from this time period were further resolved or expanded to include more isolates than previously identified through PFGE. While WGS-based analysis offers many benefits, these technologies will transform the current surveillance systems for *Salmonella*, not only in Canada but around the world. However, these transformations should not be used to diminish the importance of the various surveillance programs, their mandates, and the need for multiple streams of evidence.

Acknowledgements

I would first like to thank my supervisor Dr. Celine Nadon for not only providing me with the opportunity to complete a Master's degree, but also for her support, expertise, and compassion during my time as a graduate student. The completion of this thesis would not have been possible without her invaluable assistance throughout this journey. I would also like to thank my committee members, Dr. Chrystal Berry, Dr. Morag Graham, Dr. Claudia Narváez-Bravo, and Dr. John Wylie for proving excellent advice and guidance during this project.

I must also express my sincere gratitude to Sara Christianson, Lorelee Tschetter, and Aleisha Reimer and the teams they lead as heads of Reference Services, PulseNet Canada, and Public Health Genomics, respectively. Together you have all provided me with invaluable assistance, advice, and support during my time as a graduate student, your expertise was always appreciated. Special thanks must go out to Matthew Walker for his work on all the DNA preparations, without him this thesis would not have been completed. I must also thank the Genomics and the Bioinformatics Cores at the NML for their help and support in sequencing and data analysis throughout this project. As well, I must also acknowledge the work of the SISTR development team, including but not limited to Catherine Yoshida, James Robertson, Peter Kruczkiewicz, and Dr. Eduardo Taboada for the expertise and guidance they provided. Thanks also to all the members of the PulseNet Canada Steering Committee for their support and providing me with isolates and a project to complete.

Lastly, I must thank my parents, Dean and Josie, and my partner Meaghan for their endless patience, support, and understanding as I pursue my many goals.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

BLAST	Basic Local Alignment Search Tool
bp	Base-pair
CFIA	Canadian Food Inspection Agency
cgMLST	Core Genome Multilocus Sequence Typing
DALYs	Disability Adjust Life Years
eBG	eBurst Group
FIORP	Foodborne Illness Outbreak Response Protocol
FLASH	Fast Length Adjustment of Short Reads
hqSNV	High Quality Single Nucleotide Variant
IRIDA	Integrated Rapid Infectious Disease Analysis
LPS	Lipopolysaccharide
LB	Luria Bertani
MLST	Multilocus Sequence Typing
MLVA	Multiple Locus Variant Number Tandem Repeat Analysis
MSA	Multiple Sequence Alignment
NCBI	National Center for Biotechnology Information
NESP	National Enteric Surveillance Program
NGS	Next Generation Sequencing
NML	National Microbiology Laboratory
NTS	Non-Typhoidal *Salmonella*
PFGE	Pulsed-Field Gel Electrophoresis
PHAC	Public Health Agency of Canada
PHE	Public Health England
PMPBJ *fljB*+	Phenotypically Monophasic Paratyphi B var. Java isolates that possess the *fljB* gene
PMT *fljB*+	Phenotypically Monophasic Typhimurium isolates that possess the *fljB* gene
SGSA	*Salmonella* Genoserotyping Assay
SISTR	*Salmonella in silico* Typing Resource
SNV	Single Nucleotide Variant
SNVPhyl	Single Nucleotide Variant Phylogenomics
SPAdes	St. Petersburg Genome Assembler

SPI	*Salmonella* Pathogenicity Island
SRST2	Short Read Sequence Typing 2
ssp I	*Salmonella enterica* subspecies *enterica*
ssp II	*Salmonella enterica* subspecies *salamae*
ssp IIIa	*Salmonella enterica* subspecies *arizonae*
ssp IIIb	*Salmonella enterica* subspecies *diarizonae*
ssp IV	*Salmonella enterica* subspecies *houtenae*
ssp VI	*Salmonella enterica* subspecies *indica*
ST	Sequence Type
T3SS	Type 3 Secretion System
UPGMA	Unweighted Pair Group Method with Arithmetic Mean
wgMLST	Whole Genome Multilocus Sequence Typing
WGS	Whole Genome Sequencing
WHO	World Health Organization
WHO-Slam	WHO Collaborating Center for Reference and Research on *Salmonella*
WKL	White-Kauffmann-Le Minor

Chapter 1: Introduction

1.1 Current State of Foodborne Illnesses

Foodborne illnesses are a major burden to public health; affecting both the developed and developing world (1). These diseases are transmitted through the consumption of food contaminated with various microbial, chemical, or physical agents that are known to cause illness (1, 2). While foodborne diseases may often present as acute, mild, and self-limiting gastrointestinal syndromes, they can also lead to a variety of serious chronic health conditions or even death (1). Understanding the true burden of foodborne illnesses is an important task for public health officials and policymakers worldwide allowing for both the proper allocation of resources to address this issue and to ultimately assess the effectiveness of various intervention strategies (1, 2). However, estimating the burden of these illnesses is a complex task (1, 3). Cases of illness captured by surveillance systems are under-representative of the true burden due to under-diagnosis (care is not sought, samples not tested, agent remains unfound) and under-reporting (lack of reporting to the proper program). Estimates must account for the under-ascertainment at each step in the process of going from an individual developing an illness to that illness being reported and included in surveillance programs. Not only do these multiplicative factors vary from pathogen to pathogen, but global estimates are further hampered by regional disparity and gaps in surveillance capacity that further compound these issues (3).

In 2006, the World Health Organization (WHO) launched the Foodborne Disease Burden Epidemiology Reference Group to estimate the global burden of foodborne illnesses for 31 global hazards. Their 2015 report estimated that these hazards were responsible for 600 million cases of foodborne illness and 420,000 deaths in 2010 with a considerable portion of deaths occurring in children under five years of age (1). Disability Adjusted Life Years (DALYs) provide a way to quantify the mortality, morbidity, and disability associated with an illness and

can be thought of the number of 'healthy' years lost across a population due to a disease (4). The WHO estimates that these 31 global foodborne hazards account for 33 million DALYs worldwide in 2010 (1).

In Canada, a recent study from the Public Health Agency of Canada (PHAC) estimated that there are four million cases of domestically acquired foodborne illnesses each year (3). While the majority of these cases would be mild to moderate gastroenteritis, PHAC estimates that in Canada foodborne illnesses result in 11,600 hospitalizations and 238 deaths each year (5). Canadian cost analysis studies have not been completed on a national level (3, 5), but studies out of Ontario (6) and British Columbia (7) have estimated the cost of a gastroenteritis case to be $1089 and $1342.57, respectively, with the majority of the costs associated with loss of work. While generalizations could be drawn from these numbers it is important to note that both of these studies did not take into account the costs associated with any extraintestinal infections, sequelae, or death; therefore, the true costs of foodborne illness in Canada is likely much higher (6, 7). In fact, the 2008 outbreak of *Listeria monocytogenes* in Canada, that resulted in 57 illnesses and 24 deaths, showed just that, with an estimated cost of $2.8 million dollars per case. In total, the entire outbreak investigation was estimated to cost almost $242 million (8), indicating that the costs associated with both foodborne illnesses can be massive.

1.1.1 Current Challenges and Trends in Food Safety

Foodborne pathogens are monitored along the farm-to-fork continuum and improvements to hygiene and food safety have been incorporated both before and after harvest in attempts to reduce the burden of disease. These interventions include the vaccination of breeder and layer poultry flocks against *Salmonella enterica* PT4 in Europe, as well as targeted and improved hygiene and slaughtering practices to reduce *Escherichia coli* O157 in meats. While these

interventions have reduced the burden of these organisms within these commodities (2), evidence of subsequent reductions in human disease is limited, and further complicated by the variable collection of baseline surveillance data. In some cases, such as for *S. enterica* PT4 in Europe, the intervention strategies using enhanced hygiene, culling of infected flocks, and vaccination has controlled this pathogen but incidence of other organisms have increased as they take advantage of the recently vacated niche (2).

Global trends towards large-scale food production and the widespread distribution of products pose additional threats to food safety. Improved transportation logistics and global supply chains allow for potentially hazardous commodities to be spread internationally and incorporated into many downstream products. This coupled with the increases in the number of high-risk individuals, specifically the elderly and immunocompromised, create the potential for future large-scale outbreaks that cross geopolitical lines. As well, changing farming practices, climate change, and pathogen evolution have been postulated to lead to changes in the risk of known foodborne agents or lead to the emergence of novel zoonotic agents that could be spread through food (2).

Bacterial agents are the most monitored and understood causes of foodborne diseases due to their relative ease of detection and long history of study (2). In Canada, *Salmonella* spp., *Campylobacter* spp., *E. coli* O157, and *Listeria monocytogenes* are all considered to be priority pathogens due to high incidence rates (3), potential to cause foodborne outbreaks (2), and association with foodborne illness hospitalizations or deaths (5).

1.2 *Salmonella* Species

Salmonella spp. are Gram-negative, facultatively anaerobic bacilli within the family *Enterobacteriaceae*. They are biochemically grouped in this family by their ability to use citrate

and lysine as a sole carbon and nitrogen source, respectively; as well as their production of hydrogen sulfide on triple sugar iron agar (9). *Salmonella* spp. are responsible for an estimated 87,797 foodborne infections each year in Canada and is a leading causative agent of foodborne disease along with Norovirus, *Clostridium perfringens*, and *Campylobacter* spp., which together account for 90% of pathogen-specified illnesses in Canada (3). *Salmonella* are also a key domestically acquired foodborne pathogen associated with hospitalizations and deaths in Canada. It is estimated that almost 1000 hospitalization and 20 deaths can be attributable to domestically acquired *Salmonella* infections each year (5). *Salmonella* are also associated with a wide range of food commodities and are frequently associated with outbreaks of disease (10), making them an important priority foodborne pathogen to study.

1.2.1 *Salmonella* Nomenclature and Taxonomy

The nomenclature and taxonomy of *Salmonella* spp. has gone through numerous iterations and been the subject of much debate (11). Historically, the genus *Salmonella* was separated into many species and subspecies that were differentiated on the basis of clinical, antigenic, and biochemical properties. In 1987, using information on DNA hybridization, Le Minor and Popoff proposed a single *Salmonella* spp. that could be subdivided into seven subspecies and further divided into numerous serovars (12). Although Le Minor and Popoff's system was not formally accepted at the time, it gained widespread use and for many years there were two competing systems of nomenclature used to describe *Salmonella.* In 2005 the International Committee on Systematics of Prokaryotes issued an official update of the nomenclature scheme for *Salmonella* that took into account the taxonomic, biochemical and genetic characteristics of this group while unifying the various nomenclature systems (11).

It is now officially recognized that there are two species within the genus *Salmonella, Salmonella enterica* and *Salmonella bongori. S. enterica* can be further subdivided into six subspecies: *S. enterica* subsp. *enterica* (I)*, S. enterica* subsp. *salamae* (II), *S. enterica* subsp. *arizonae* (IIIa), *S. enterica* subsp. *diarizonae* (IIIb)*, S. enterica* subsp. *houtenae* (IV)*,* and *S. enterica* subsp. *indica* (VI) (9, 11, 13). These subspecies are differentiated on the basis of biochemical reactions (9), and can further be subdivided into serotypes based on their interactions with antiserum (13). Of note, only *S. enterica* subsp. *enterica* are considered pathogenic to warm-blooded animals (9) and make up 99.5 percent of all clinical *Salmonella* isolates encountered (13).

1.2.2 Clinical Manifestations

Salmonella infections can be broadly categorized into typhoidal or non-typhoidal infections (9, 14). Typhoidal infections or typhoid fever is a serious bloodstream infection that is not often associated with foodborne transmission in North America but is a major problem in other parts of the world, specifically Africa, the Middle East, and South East Asia. In fact, typhoidal *Salmonella* infections are estimated to be the most common cause of non-diarrheal foodborne illness related deaths in the world (1). Typhoidal infections are the result of a certain subtype of invasive *Salmonella*: specifically isolates whose only known reservoir is humans, and include the serovars Typhi, Sendai, and Paratyphi A, B, or C (14). These infections have a long incubation period (9, 14) and are denoted by a sustained and high fever, abdominal pain, hepatosplenomegaly (swelling of liver and spleen), rash, and dry cough. Typhoidal *Salmonella* do not induce intestinal inflammation but instead gain access to the underlying lymphatic tissues, multiplying intracellularly, leading to the systemic infection. Up to 10% of untreated typhoidal patients can shed the bacteria in their feces long after the infection has been cleared, and others

may become asymptomatic and chronic carriers of the bacteria. These chronic carriers are believed to be important in the disease transmission of *Salmonella* Typhi as illustrated by the famous case of Typhoid Mary, an early 20^{th}-century cook who is believed to have infected up to 54 people in the New York City area. Typhoidal infections require immediate treatment with antibiotics, specifically the fluoroquinolones and third generation cephalosporins (14).

Non-typhoidal *Salmonella* (NTS) meanwhile are the more frequent source of foodborne *Salmonella* infections in North America (1) and infections with NTS are often of intestinal origin with symptoms such as diarrhea, fever, and abdominal cramps. NTS can also lead to invasive infections, most often in immunocompromised patients with symptoms closely resembling typhoidal infections. Certain NTS serovars are more likely to be associated with invasive infections, although the exact virulence factors responsible are not well understood. Treatment for NTS consists of rehydration and electrolyte balance as the gastrointestinal illness is often self-limiting and clears up within a week. For severe cases of gastroenteritis or invasive infections, antimicrobials may be required, but are otherwise contraindicated due to the risk of potentially prolonging the illness (14).

1.2.3 Pathogenesis

The infectious dose of *Salmonella* ranges from as few as 30 to as many as 10^9 colony forming units and is associated not only with variability in immunological host factors and bacterial virulence, but could also be influenced by the composition of the contaminated food (15, 16). High-fat content has been shown to increase the thermal resistance of *Salmonella*, and presumably, fats and other components could protect the bacterium from gastric juices (17). Genomic studies of NTS have revealed the presence of five conserved *Salmonella* pathogenicity islands (SPI) located on the chromosome. These SPIs are unique clusters of conserved genes that

are essential to the organism's virulence and appear to have been acquired through horizontal gene transfer. The major virulence factor encoded by SPIs are two separate type III secretion systems (T3SS) (15, 16). The two T3SS are expressed at different times in the infection, with one being involved in the attachment and invasion of intestinal cells and the other for survival within macrophages (18). Meanwhile, genomic studies of typhoidal *Salmonella* serovars have shown the functional disruption or inactivation of many genes, including multiple effectors of the T3SS. As well, typhoidal *Salmonella* possess further SPIs that are not seen in NTS. It is believed that these changes lead to the unique disease outcomes of the typhoidal serovars and result in the reduced intestinal inflammatory response caused by these isolates (14).

Diarrhea seen with NTS infections is partially the result of the induction of an inflammatory response by the host immune system (15, 16, 18). The T3SS binds to the host intestinal cells and translocates effector proteins into the cytoplasm. These effector proteins result in cytoskeletal rearrangements; thereby, mediating the invasion of the bacteria into the cell, while also disrupting the tight junctions between intestinal cells (18). At this point, host polymorphonuclear leukocytes migrate into the intestinal lumen to combat the invading bacteria and induce an inflammatory response leading to diarrhea (15, 18). It is believed that NTS uses the inflammation of the gut to outcompete the normal flora of the host (14).

Salmonella also produces a diarrheagenic enterotoxin that is expressed within hours of attachment. The expression of this toxin in the cytoplasm of infected intestinal cells leads to large increases in the concentration of cyclic AMP in the cell. The rise in cyclic AMP forces Cl^- ions out of the cell and blocks the uptake of Na^+ ions by the cell. The resulting electrolyte imbalance between the cell and the intestinal lumen induces the movement of water from the cell into the intestinal lumen and is also believed to play a role in the development of diarrhea (15).

1.2.4 Transmission of *Salmonella* in Food Production

Salmonella have shown the ability to resist numerous stresses presenting a major challenge to food safety. Specifically, *Salmonella* has been shown to: survive for long periods of time under dry conditions and in low moisture foods; display high thermal tolerance, especially when subjected to other sub-lethal stresses; and survive in acidic environments (15, 17). However, cross contamination of products, due to poor sanitation, facility/equipment design, manufacturing practices, ingredient control, and pest control, remains the most significant risk for the presence of *Salmonella* in food products (17). *Salmonella* infections are most often associated with contaminated poultry products, including eggs; however, the bacteria can also be spread through dairy, fresh produce, or other processed and ready-to-eat foods (9, 15, 16).

In the United States, microbial contamination of food commodities is the most significant reason for a food recall event and the majority of these involve *Salmonella*. Over a nine-year period, microbial contamination was responsible for 1,395 recalls in the United States, with *Salmonella* being involved in 71% of these (19). Unfortunately, formal data on this is lacking in Canada. However, examination of the 2015 and 2016 recall lists that triggered public warnings from the Canadian Food Inspection Agency (CFIA) revealed that microbial contamination is responsible for just under half of all public recalls, with *Salmonella* being responsible for almost 30% of these, second only to *Clostridium botulinum* (http://www.inspection.gc.ca/about-the-cfia/newsroom/food-recall-warnings/complete-listing/eng/1351519587174/1351519588221).

1.2.5 Outbreaks of *Salmonella* in Food

An analysis of *Salmonella* associated outbreaks in the United States between 1998-2008 identified a total of 403 outbreaks that were linked to a single food commodity. Overall eggs, chicken, pork, fruit and turkey were the top five commodities, responsible for more than 25

outbreaks each (10). Although a large scale analysis has not been performed yet in Canada, the examination of outbreaks identified by PulseNet Canada has revealed that past Canadian outbreaks have been linked to chicken, produce, and various processed and prepared foods (data not shown). It has also been noted that certain types of *Salmonella* outbreaks are more commonly attributed to various food commodities, indicating their preferred host or natural reservoirs. For example, outbreaks of *Salmonella* serovar Enteritidis, whose natural reservoir is chicken, are more likely to be associated with poultry products; while *Salmonella* serovar Weltevreden outbreaks are more likely to be linked to seafood products, as its natural reservoir is considered aquatic animals. It is important to note that food commodities of sporadic infections may possibly differ from those linked to outbreaks as there has been little data collected on food commodities linked to sporadic infections (10).

Investigations into the microbial contamination of food can become extremely complex. Ingredient-driven contamination events are a major source of concern as a single contaminated ingredient can be used in many downstream products and confound any outbreak investigation (19, 20). For example, a 2008-2009 international outbreak of 530 cases of salmonellosis was linked to the consumption of peanut butter and peanut butter containing products. In this example, epidemiological and laboratory investigations linked the outbreak strain to King Nut creamy peanut butter, produced at a facility in Georgia. Peanut paste that was produced in the same plant was also found to be contaminated with the outbreak strain through laboratory testing. The contaminated peanut paste was sold and distributed to companies in over 23 countries for use in other peanut-butter containing products, further complicating the outbreak. Recalls of the peanut butter and products containing peanuts that were produced in the plant ultimately resulted in the recall of over 430 products from 54 companies by January 2009 (20).

Microbial contamination of food can also be complicated further by the presence of multiple diverse strains or serotypes within a single commodity. While these types of outbreaks are not commonly reported, it is believed that they might occur more frequently than is realized, as their detection is limited (21). Multiple strains or serovars within a single commodity can pose significant challenges to the investigation. Diverse molecular subtyping patterns would not be linked through laboratory investigations due to the large number of differences (22). This could also cloud epidemiological investigatory signals by reducing the number of cases linked together; thereby reducing the power to implicate a single source. Past detection of multi-strain outbreaks have required the screening of multiple single colonies from co-infected individuals (21), or the detection of other disease-causing isolates from an already implicated product (23).

1.3 Overview of Surveillance for *Salmonella* in Canada

Surveillance of *Salmonella* in Canada is carried out along the farm to fork to human disease continuum via multiple surveillance programs. The information collected from these programs can be integrated together using a One Health model for surveillance, which is a multidisciplinary approach that combines and recognizes the interplay between human, animal, economic, and environmental systems. This internationally accepted model has been used to provide a deeper understanding of *Salmonella* in Canada and draws heavily from the various national surveillance programs (24). On a broad level, programs such as FoodNet Canada use a sentinel site model to survey both the incidence of disease in the community and the sources attributed to those diseases. Its main objectives surround understanding the broad trends in enteric disease burden, potential sources of disease within a community, and the strength of various intervention strategies (25). In real time, human enteric disease is specifically captured by two programs in Canada, the National Enteric Surveillance Program (NESP) and PulseNet

Canada (24). Information from these two programs is used to provide the clinical microbiological evidence on multi-jurisdictional outbreaks of enteric illnesses as described in Canada's Foodborne Illness Outbreak Response Protocol (FIORP) (26).

1.3.1 National Enteric Surveillance Program (NESP)

NESP provides the rapid and weekly analysis of the etiological agents causing enteric illnesses in Canada providing an early warning signal to potential outbreaks. Data collected by NESP is at the lowest levels of microbial resolution, specifically focusing on the species, serovar, and phage-type of enteric disease-causing isolates. Analysis of the compiled data involves comparing current levels of enteric diseases with the 5-year retrospective median value of cases determined from the long standing historical records maintained by NESP. This type of analysis can be completed using Poisson statistics, which allows for the quick detection of small shifts away from mean values. A significant increase in the number of reported cases of an enteric illness is used to alert NESP stakeholders to potential disease clusters and outbreaks. Within the larger surveillance system for enteric diseases, NESP data provides the timeliest and most basic level of classification for enteric illnesses and the information collected by NESP is integrated into and informs the downstream surveillance activities of other programs such as PulseNet Canada (27).

Salmonella remains the most commonly reported enteric pathogen to NESP, with a total of 7,851 isolates reported to NESP in 2014. Between 2009-2014, the crude incidence rate of reported *Salmonella* infections (cases per 100,000 persons) has remained fairly consistent with an average of 19.8 cases per 100,000 persons and fluctuating between a low of 17.8 in 2013 to a high of 22.0 in 2014 (28). It is important to note that a calculation of reported incidence rates of enteric pathogens using data collected by NESP are not necessarily equivalent to the true

incidence rates, as these calculated rates are reliant on and reflect the testing and reporting practices within provinces. However, for some organisms, such as for *Salmonella*, these rates are considered to be more in line with the true incidence rates due to the compliance with routine testing and forwarding of information to NESP (27) .

1.3.2 PulseNet Canada

PulseNet Canada is the national, standardized, high resolution, molecular subtyping network for bacterial enteric pathogen surveillance. In real time PulseNet Canada performs and analyzes high-resolution molecular subtyping techniques on priority enteric pathogens for the identification and investigation of foodborne disease outbreaks. PulseNet Canada is based on an internationally developed and widely used model, with networks serving the United States, Europe, Latin America and the Caribbean, Africa, the Middle East, and the Asia Pacific regions. Together these networks use standardized methodologies to track foodborne infections within their respected regions and share information of importance within and between the regions (29).

In Canada, participating laboratories, both provincial/territorial and federal, upload molecular subtyping patterns to online databases for the rapid detection of clusters of disease (29). Isolates with matching molecular subtyping profiles are identified as clusters of disease (29, 30), which are reported back to the network of participants through the online PulseNet Canada forums (29). Epidemiologists can then confirm the presence of a true outbreak through epidemiological investigations (30), at which time an outbreak investigation could be activated through FIORP (26). The molecular subtyping performed by PulseNet Canada allows for the rapid identification of clusters of disease that may be dispersed either geographically or temporally. This information is valuable to outbreak detection as illnesses are increasingly spread across multiple jurisdictions, due to the wide distribution networks of food (29, 31).

The PulseNet model for bacterial pathogen surveillance has had major success in both Canada (29, 32-34) and around the globe (29, 35, 36). The strength of the PulseNet model comes from the rapid analysis of molecular subtyping data, the use of highly standardized protocols and techniques allowing for interlaboratory comparisons, and the enhanced levels of communication between partners that allow for the linkage of what may have appeared as sporadic cases of illness to a well-defined cluster (29). An economic and health benefit analysis by PulseNet USA found that the program was directly responsible for a reduction of over 19,800 cases of illness in a single year through the initiation of recalls based on the early identification of outbreaks and source attribution. As well, the authors also estimated that PulseNet USA had reduced foodborne illnesses by over 267,000 cases each year, due to indirect effects from their work. These indirect effects include providing information to regulators and industry from their investigations to better craft food safety practices, as well as providing incentives to implement these improved practices. The reduction of foodborne illnesses in the USA due to PulseNet USA's activities are estimated to have led to a cost savings of $544 million per year, mostly from the medical and lost productivity costs of illness (37). While a similar large scale economic and health benefit analysis has not yet been conducted for PulseNet Canada, its work has been instrumental in detecting, solving, and preventing outbreaks of foodborne illnesses within Canada (29, 32-34).

1.4 Methods for the Identification and Subtyping of *Salmonella*

Traditionally *Salmonella* is first identified from fecal or environmental specimens using differential plating media. These differential media, such as MacConkey's agar, may allow for the growth of certain enteric pathogens, including *Salmonella,* or may be highly selective for *Salmonella* alone, such as the case of bismuth sulfite agar. Ultimately, a presumed *Salmonella* isolate will undergo a series of biochemical tests to screen for and confirm the organism. These

tests can include the triple sugar iron slants, in which *Salmonella* produces a red slant, yellow butt, black precipitate, and hydrogen sulfide gas, highlighting the ability of *Salmonella* to ferment glucose and reduce sulfur. Other tests such as the Simmons citrate agar test display the ability of *Salmonella* to use citrate as its sole carbon source, while the urease test shows it is unable to produce urease (9). Following identification, multiple methods for the enhanced subtyping and classification of *Salmonella* exist at varying levels of subtyping resolution (38). The results from various subtyping methodologies are collected by the different surveillance programs in Canada. The generation and collection of this information is based on both the tests resolution and the individual program mandate (24). NESP collects information on *Salmonella* using the low-resolution tests, serotyping and phage typing (27), while PulseNet Canada collects information from high-resolution tests, such as pulsed-field gel electrophoresis (PFGE)(29) and multiple locus variant tandem repeat analysis (MLVA) (30). Other molecular subtyping tests of varying resolution such as multilocus sequence typing (MLST) may also be performed as needed by research scientists or public health officials (38).

1.4.1 Serotyping

Serotyping classifies *Salmonella* isolates on the basis of the cell surface antigens, specifically the somatic O- and the two variably expressed flagellar H-antigens, denoted as H1 and H2 (39-42). The O-antigen is the outermost carbohydrate component of the Gram-negative lipopolysaccharide (LPS) (40, 43). The production of the O-antigen is the result of a complex molecular pathway and the genes responsible for the biosynthesis, assembly, and transport of LPS is encoded by the *rfb* operon in *Salmonella*. This operon shows low GC content and is thought to have been acquired through lateral gene transfer (44). Several O-antigens may be expressed at the surface of the cell (13), and some are expressed through the conversion of a

bacteriophage and are therefore denoted in the antigenic formula by an underline (13, 45). The H-antigen is considered to be a diphasic antigen as most *Salmonella* possess two different genes (*fliC* and *fljB*) its production (40, 46). Interestingly, a single *Salmonella* cell will only express one H-antigen in its life, but both will still be present in a pure culture via a process called flagellar phase variation. This process occurs during chromosomal replication via the reversible inversion of a promoter that controls the transcription of the phase 2 gene, *fljB,* and the phase 1 inhibitor gene, *fljA.* When this promoter is properly oriented both the H2 antigen and inhibitor protein FljA are produced. FljA will repress the translation of the H1 antigen by rapidly degrading *fliC* mRNA, and only the H2 antigen will be expressed. Once every 1,000 to 10,0000 cell generations, the promoter region for *fliB* and *fliA* is recombined leading to the loss of the H2 antigen and FljA; thereby, allowing for the translation of *fliC* and the expression of the H1 antigen (46).

A total of 64 O- and 114 H- antigens have been identified (40), which can be combined into 2,579 recognized serovars (13) each with an antigenic formula reported as O:H1:H2 (13, 40, 43). The original scheme for serotyping *Salmonella*, the Kauffmann-White scheme, was published in 1934 and identified just 44 serovars; in the years since, the number of named serovars has jumped dramatically, mainly due to the work of Leon Le Minor. Currently, the WHO Collaborating Center for Reference and Research on *Salmonella* (WHO-Salm) updates and maintains the serotyping scheme for this genus and has renamed the scheme to the White-Kauffmann-Le Minor (WKL) scheme, to reflect Le Minor's work (13). While there are thousands of identifiable serovars, it is important to note that the majority of human clinical disease is the result of a select few important serovars (47). For example, in Canada, the 20 serovars most commonly linked to human disease represent about 85% of all *Salmonella* disease

within a given year, with the top serovar, Enteriditis, responsible for almost half of all reported human infections (27) .

Historically many serovars of *Salmonella* were considered separate species and were given names that denoted important information about the serovar (11, 13). Some were named for the condition they produced, such as serovar Typhi that causes typhoid fever. Others included host specificity in the name, such as serovar Aborusovis that induces abortion in sheep. Occasionally the host specificity and condition initially described was not correct, such as the case of Typhimurium. This serovar was first identified to cause a typhoid-like syndrome in mice and was later found to be a major cause of gastroenteritis in most mammals including humans. To prevent confusion, geographical origins of novel serotypes were used later on leading to serovars such as Heidelberg, Montevideo, and Sandiego. However, with the ever-expanding list of serovars and the improved taxonomy of the genus, the scheme evolved and names are now only kept for *S. enterica* subsp, *enterica,* while serovars from all other subspecies are denoted by their subspecies Roman numeral and their antigenic formula. Since the serotype name is not a formal taxonomic designation it is not italicized but instead, is capitalized (13).

Traditionally, serotyping is performed through the phenotypic characterization of the O- and H-antigens of *Salmonella* via the slide agglutination test, in which the clumping of cells is observed in response to specific antisera. Although this technique is widely used (40) it can be time-consuming and laborious requiring highly trained technicians as well as the maintenance, storage, and quality control of over 150 antisera (39, 40, 42). As well, five to eight percent of all *Salmonella* isolates remain untypeable via this method (39). For these reasons, there are few reference laboratories with the capacity to classify all *Salmonella* serovars (48).

Molecular serotyping platforms have also been developed and evaluated in more recent years and include both PCR and array-based techniques. These methodologies allow for the detection and characterization of the antigenic markers using a molecular method and represent a major advantage due to their high-throughput capabilities and the production of results in a timely manner (40, 49-51). One PCR-based serotyping methodology utilized three multiplex PCR reactions one each to identify the genes responsible for the H1, H2, and O serogroup. While this method was able to identify 84.9% of the isolates tested in its initial validation, it is important to note that correct results were reported for only 15 of the 33 serovars included in the validation (51). DNA microarray methodologies such as the PremiTest assay or the *Salmonella* genoserotyping assay (SGSA) have been developed to identify a multitude of serovars (49, 50). The PermiTest targets non-antigen genomic markers that have been identified to be unique to specific serovars (40, 49), and has been shown to identify 94.7% of the *Salmonella* isolates reported to the Belgian reference laboratory over a nine month period (49). Newer versions of the test are increasing the genetic markers used and the number of different serovars it is capable of identifying (40). The SGSA meanwhile targets the genes responsible for antigen production and has been shown to rapidly and successfully identify 57 of the most commonly reported serovars (50). Ultimately while many of these methodologies have been developed and proposed few have been widely adopted, in part due to the limited number of identifiable serovars from these methods (40).

1.4.2 Phage typing

Some serovars can be further classified phenotypically via phage typing. This classification scheme discriminates isolates within a serovar on the ability to be lysed by certain bacteriophages (45). The number and types of phages used in the typing schemes differ between

serovars; however, the overall technique remains the same. A specialized agar plate is inoculated for confluent growth onto which the typing phages are dropped on the surface. The size, number, and degree of cellular lysis are recorded for each typing phage and an overall phage type is determined using a schema (45). Phage typing was an important technique for subtyping of *Salmonella* Enteritidis and Heidelberg isolates, as other routine methods such as PFGE did not provide optimal discrimination (52, 53). Phage typing could also provide information on the geographical origin of *Salmonella* Enteritidis isolates. For example, PT8, PT13a, and PT13 were the most dominant phage types in North America (27, 52, 54) while PT4, PT1, PT6a and PT14b all represent dominant phage types in other parts of the world (52, 54). The detection of these phage types in North America indicates recent travel and their incidence would peak during major travel seasons (52).

The usefulness of phage typing has been called into question (54, 55). Although the procedures are of relative ease, the interpretation of results is subjective and can lead to different phage type designations between laboratories, creating confusion in outbreak investigations. For example, a 2003 outbreak of *Salmonella* Typhimurium phage type DT108 in Sweden was linked to Danish pork meat, yet this definitive type had not been detected in Denmark and the pork meat was thereby ruled out. Further investigation then discovered that the Typhimurium DT108 phage type in Sweden had been classified as Typhimurium DT170 in Denmark eventually resulting in the Danish pork meat being ruled as the probable source. These types of subjective interpretations can have a big impact on the timely recall of products and can greatly impact public health (55). As well, phage types have been shown to have weak phylogenetic relationships, calling into question any use for inferring phylogenetic information. Plasmids have been shown to convert *Salmonella* Enteritidis isolates between phage types, potentially

explaining the weak phylogenetic information provided by phage typing (54). Recently the Public Health England (PHE) laboratory responsible for the production of phages for this test has stopped production, cementing the fate of phage typing as solely a historical test (Personal Communication, Celine Nadon, 2017).

1.4.3 Pulsed-Field Gel Electrophoresis

PFGE has been considered the gold standard technique for the molecular subtyping of enteric organisms for two decades and is the basic molecular subtyping technique of the PulseNet International surveillance model (29). This technique relies on the enzymatic fragmentation of the bacterial chromosome followed by a process of separation and visualization to produce a characteristic banding pattern used for comparison. The interlaboratory reproducibility of PFGE can be challenging, as there are many areas for potential variation including in the sample preparation, fragment separation and visualization steps (38, 56, 57). To counteract these factors, PulseNet International has put into place highly standardized protocols for performing PFGE and analyzing the PFGE banding patterns, as well as mandating the certification and annual proficiency of all participants. This allows for proper interlaboratory reproducibility; thereby, giving this technique epidemiological significance (31, 38, 56, 58).

While the individual PFGE protocols for the various organisms differ, they do have overarching similarities. The bacterial cell is first suspended in an agarose plug for stabilization. The isolation of the entire bacterial chromosome occurs in the agarose plug through a series of steps involving cellular lysis and washing to remove the unwanted components. The chromosome is then subjected to a restriction enzyme that will cut or fragment the DNA at a specific sequence of nucleotide bases, called restriction sites. The frequency of the specified restriction sites can be estimated, and various restriction enzymes are used in PFGE depending

on the organism under study (38, 57). The optimal restriction enzyme for PFGE will produce 10 to 25 fragments of DNA ranging in size from 20kb to greater than 1MB (38). For *Salmonella* PulseNet International protocols use *Xba*I and *Bln*I as the primary and secondary restriction enzymes, respectively (59). The fragmented DNA is then separated via an agarose gel and an electrical current. Due to the size of the restriction fragments produced, the electrical current is periodically reoriented along a 120° axis. Smaller DNA fragments are able to reorient faster within the gel matrix and the equidistant migration from left to right results in the separation of the DNA fragments in straight lines (38, 57). To further standardize the results between labs, a size standard, the *Xba*I digestion of *Salmonella* Braenderup strain H9812, is run in every fourth to fifth lane on all PulseNet International PFGE gels. This size standard contains 15 evenly distributed bands allowing for consistent normalization between gels and the comparison across both geographical and temporal distances (56).

The characteristic banding patterns produced by PFGE are compared visually (57). Due to the large numbers of PFGE patterns produced for PulseNet programs, the comparison of patterns is done through computer-assisted methods and technicians then confirm the analysis (56). The comparison criterion differs between organisms and is based on the underlying genetic variability present within that organism (57). The Tenover criteria were the first guidelines for outbreak investigations of nosocomial infections using PFGE profile data. The criteria state that profiles of closely related isolates can differ by up to three bands, and potentially related isolates can differ by up to six bands. Tenover reasoned that a single genetic event at the restriction site would create a difference of three bands between two isolates, as a loss of a restriction site would merge two smaller fragments into one larger fragment for an isolate. The loss of the two smaller fragments and gain of the larger one would create the three band difference between the isolates

(60). Although these criteria have been highly useful, they are not applicable to all situations, especially foodborne pathogens such as *Salmonella*. These criteria assume all fragments from a digest are visible in the gel and that plasmid content of the isolates is stable, both of which are not true in foodborne pathogens. As well, since the restriction sites of PFGE enzymes are rare, only a tiny fraction of mutations would likely occur at these sites, meaning several events could have occurred before any pattern is altered (22).

PulseNet Canada protocols frequently require the secondary enzymatic digestion using *Bln*I on separate plugs to confirm the results or provide further discrimination (56). Exact PFGE matches, or highly similar PFGE patterns, may be considered a cluster of isolates and these isolates are potentially part of an outbreak. The interpretation guidelines for PFGE data generated by PulseNet Canada were adapted from the recommendations published in a 2006 study from Barrett *et al.* on interpreting PFGE results from foodborne pathogens (22). The specific guidelines utilized by PulseNet Canada are outlined in a document that was prepared by Health Canada, PHAC, and CFIA which outlines the interpretation criteria to be utilized for all evidence that is gathered during a foodborne illness outbreak investigation. The strength of PFGE pattern matches are assessed on four categories, specifically: (a) the diversity of the organism/serotype; (b) distinguishability of patterns; (c) frequency of pattern combination; and (d) availability of alternative subtyping data (Table 1) (61). Based on the strength of the laboratory evidence collected, an outbreak investigation may occur, as outlined in FIORP (26, 56, 58, 61).

Table 1: Interpretation guidelines for determining the strength of isolate matches by PFGE data in the context of a foodborne illness outbreak. Reproduced from data presented in Canada's Weight of Evidence document (61).

Criteria	Nature of Evidence	Weight
A. Diversity of organism/serotype[1]	Organism/Serotype shows diverse PFGE patterns among previous sporadic cases.	Strong
	Little to no historical data exists for this organism/serotype.	
	Organism/Serotype shows little diversity by PFGE among previous sporadic cases.	Weak
B. Distinguishability of pattern combination	Isolates indistinguishable by two enzymes.	Strong
	Isolates are indistinguishable by first enzyme but distinguishable by second enzyme; with only minor differences detected at the low molecular weight region.	
	Isolates are distinguishable by first and second enzyme, but differences are minor and at the low molecular weight region.	
	Isolates do not match; differentiated by multiple bands specifically within the high molecular weight region.	Weak
C. Frequency of pattern combination	Isolates display a new PFGE pattern or new combination of patterns.	Strong
	PFGE pattern or combination is not new but is not commonly seen.	
	PFGE pattern or combination is common.	Weak
D. Availability of alternative subtyping data	Additional data is available on the isolates (phage types, antimicrobial resistance profiles, MLVA, MLST, WGS, etc) and is in agreement with the PFGE data.	Strong
	No additional data is available on the isolates	
	Additional data is available on the isolates but is not in agreement with PFGE data.	Weak

[1]Deteremined through historical data

Although PFGE has been the gold standard for two decades, it is rapidly losing this status to newer technologies. Its longevity and widespread use within the public health community is related to many of its important characteristics and features. PFGE can be applied to almost any organism, and nearly all strains within an organism can be typed using this method. PFGE is also highly epidemiologically concordant and suitably discriminatory for many bacterial species,

grouping the epidemiologically related isolates together while excluding non-related strains (38, 57). Lastly, through the use of the highly-standardized protocols of PulseNet Canada, PFGE is highly reproducible between labs, allowing for the proper comparability required for a typing method (31, 56, 58). However, its limitations, especially in regards to its lengthy labour intensive process that is not optimized for high-throughput analysis and the lack of pattern diversity among certain organisms/serotypes have led to the supplementation of this data with alternative typing data and the pursuit of novel higher resolution techniques (38).

1.4.4 Multiple Locus Variant Tandem Repeat Analysis

MLVA is another highly useful molecular subtyping technique that subtypes organisms on the basis of repetitive DNA elements, whose copy number may differ between strains due to DNA replication errors (38, 62). The evolutionary clock speed at each locus is variable and epidemiological information can be inferred from the differences in the copy numbers that occur at multiple loci (38). MLVA utilizes fluorescently labeled multiplex PCR reactions to amplify these repeating loci. Utilizing fluorescent primers, technicians can detect multiple loci in one multiplex PCR reaction even if the loci are of overlapping sizes. The sizes of the amplicons are determined using capillary electrophoresis, giving an electropherogram output that shows the size of the bands, determined using a known standard, and their fluorescent output. The size of the individual bands is used to calculate the number of repeats at each locus and this information is converted to a standard integer number, giving the strain a MLVA profile that can be easily compared between laboratories (38, 62).

While MLVA is highly discriminatory, displays a higher throughput, and is overall less labour intensive compared to PFGE, it will likely not replace PFGE but instead complement its analysis (30, 62). MLVA lacks the universal applicability across organisms and serotypes that is

seen with PFGE; thereby preventing its wholesale replacement (38). However, many surveillance programs are supplementing PFGE data with MLVA data for highly clonal organisms due to the enhanced discrimination these two methods provide (30, 62). A retrospective analysis of *E. coli* O157 clusters in Canada found that PFGE+MLVA data analysis altered the case categorization for 60% of the clusters investigated. It was also shown that this combined data had higher concordance with the available epidemiological data than just using PFGE data alone (30). The results from this study greatly impacted the routine practice of PulseNet Canada and the combined PFGE+MLVA data analysis is now undertaken for the surveillance, cluster detection, and outbreak response of *E. coli* O157 (30), *Salmonella* Enteritidis, and *Salmonella* Typhimurium (data not shown).

1.4.5 Multilocus Sequence Typing

MLST is molecular subtyping method that was first introduced in the late 1990s (63, 64). This subtyping methodology was one of the first to apply DNA sequencing technologies to classify bacterial isolates (38). MLST works by sequencing the genome at very specific loci, most often across 350-600 base pair (bp) stretches found within five to ten highly conserved housekeeping genes (38, 64, 65). The MLST scheme for *Salmonella* specifically looks at fragments from the genes *aroC, dnaN, hemD, hisD, purE, sucA,* and *thrA* (66). Each unique sequence found at an individual locus is assigned an allele number, that was determined in the order of its discovery. Together the allele numbers assigned across all loci are used to define an individual sequence type (ST), which can then be grouped into larger clonal complexes or eBurst Groups (eBGs) of isolates that share the majority of alleles together (63, 64). Previous research has correlated numerous important phenotypic characteristics to individual ST or to larger clonal complexes (65). The unambiguous data generation and readily accessible public databases that

hold the ST information of MLST schemes have made this an incredibly successful and important subtyping system (38). However, the accumulation of mutations within housekeeping genes is relatively slow, and therefore MLST provides limited discriminatory power to properly differentiate isolates for the purposes of cluster detection and outbreak investigation (38). MLST is, however, still popular in some contexts, especially in the investigation into and understanding of population structures and evolutionary histories of organisms (66, 67).

1.5 Whole Genome Sequencing and the Genomics Revolution

Theoretically, all the relevant subtyping data about an organism is found within its genome. However, for many years the time and costs associated with sequencing an organism's entire genome made this solely a technical achievement left to researchers (68). The development of next-generation sequencing (NGS) technologies has greatly reduced the cost and time associated with sequencing. Correspondingly, the emergence of the interdisciplinary field of bioinformatics, which utilizes mathematical, statistical, and computer science based concepts to analyze biological datasets, has allowed scientists across a range of disciplines to actually take advantage of these large datasets. Together these technological advancements have placed whole genome sequencing (WGS) in a position to revolutionize surveillance laboratories (38, 68)

1.5.1 Sequencing-Based Techniques

1.5.1.1 First Generation Sequencing

First generation or the Sanger approach to sequencing utilizes a DNA polymerase that occasionally adds labeled chain terminating dideoxynucleotides to a replicating strand of DNA during replication. The DNA sequence itself can then be determined manually via gel electrophoresis by separating the products of the experiment to the resolution of single nucleotides. The radiolabel at each position could be read off and would correspond to the

nucleotide base at that position. The automated Sanger sequencing platforms were developed later and used capillary electrophoresis to separate the fragments. These platforms allowed for higher throughput of samples compared to the manual determination and were used to sequence and finish the numerous bacterial genomes as well as the first human genome. Sanger sequencing techniques are highly accurate, with long read lengths, however in comparison to next generation sequencing (NGS) technologies their throughput, even for automated determination, was lacking (69).

1.5.1.2 Next Generation Sequencing

There are numerous NGS technologies that utilize a combination of novel strategies for preparing DNA templates and determining its sequence. Together these NGS technologies further improved upon the throughput of Sanger sequencing and led to the explosion of sequencing data available for analysis. NGS technologies produce shorter sequence reads and require novel computational approaches to piece them back together (70, 71). The Illumina MiSeq is currently one of the most widely used sequencing platforms and is the sequencer of choice for PulseNet Canada laboratories. This platform can produce paired-end reads of up to 2x 300 bp in length and can rapidly sequence multiple isolates in parallel (69).

The Illumina MiSeq platform requires prepared genomic DNA for sequencing. Genomic DNA must first be fragmented, either enzymatically or mechanically. Universal adaptors are then ligated to both ends of the fragment and the primary adaptors are allowed to bind to a complementary short oligo attached to a specialized glass slide, called the flow cell. DNA polymerases then work to create a cluster of single-stranded DNA molecules through a process called bridge amplification, where other attached oligos (complementary to the secondary adaptor) are used as primers for the DNA polymerase. In the end, each fragment initially

produced is clonally amplified, forming little clusters of the single stranded template attached to the glass slide. The production of these clonally amplified fragments are crucial to the sequencing step as the imaging technology used to detect the DNA sequences cannot detect single fluorescent events (70).

The sequence of the DNA is then determined through the addition of fluorescently labeled 3' reversibly terminated nucleotides (70, 71). A DNA polymerase will add these nucleotides to the primed clonally amplified templates, and the 3' terminator will block any further nucleotide addition. Following the addition, the excess nucleotides are washed away and a fluorescent image is taken before the 3' terminating group and the fluorescent dye is cleaved and washed away. This cycle is repeated, and the corresponding fluorescent images produced are used to determine the sequence of the fragment, producing a sequence read (70). During a sequencing run, millions of sequence reads are produced in a parallel fashion on the flow cell, allowing the Illumina MiSeq platform (Illumina Inc., San Diego, CA, USA) to sequence 1.5-2 gigabases in 27 hours, meaning the raw sequence for a bacterial genome could be generated in a day (72).

1.5.1.3 Third Generation Sequencing

Third generation sequencers began to appear on the market in the late 2000s and have differentiated themselves from NGS platforms through their longer read lengths and data generation from single molecules. However, these advances in longer read lengths have also resulted in increased error rates, requiring further and new advancements in computational processes to overcome these error rates. Third generation platforms such as Oxford Nanopore MinION (Oxford Nanopore Technologies Inc., Oxford, United Kingdom) have brought WGS out of the laboratory environment and into the field, as they require minimal reagents and equipment.

However, their adoption and use by the larger microbiological community remains to be seen, and currently, most third generation sequencing technologies are seen as a complement to NGS technologies, not as a direct replacement (69).

1.5.2 Overview of WGS Data Analysis

Before any large WGS-based project can begin, such as performing surveillance activities, understanding and managing the computational requirements of the project is required. The large amount of data generated from these projects requires specialized high-end machines and data centres to store and process the data. Numerous cloud-based computing and storage systems are now commercially available to handle and process the large datasets (73). In Canada, the Integrated Rapid Infectious Disease Analysis (IRIDA) platform provides the computational and bioinformatic requirements of WGS data generated by PHAC. IRIDA allows for the sharing of WGS data to specified partners while maintaining data privacy and security. The IRIDA platform also allows for the seamless integration of bioinformatic analysis pipelines through a local instance of the web-based bioinformatics program, Galaxy. IRIDA provides users with a one-stop-shop of easy to use data management and analysis tools in a secure web-based environment (74).

The bioinformatic analysis of WGS data can be separated into three forms of analysis – primary, secondary, and tertiary. Primary analysis is concerned with the quality control procedures to ensure data generated is accurate and robust for further downstream analysis. Secondary analysis involves applications or procedures that transform the short read data into a more usable and accessible form for further downstream analysis. Lastly, tertiary analysis involves the process of placing the WGS data within the broader population context (73). The bioinformatic analysis of WGS data for large scale surveillance projects of *Salmonella* can also

be organized by the levels of resolution the data analysis provides (69). Low-level data analysis resolution can provide information on classical species designations via ultrafast taxonomic classification tools. While the use of these tools is mainly limited to metagenomics applications, these tools can also be co-opted for detecting contamination in samples (75). Meanwhile higher levels of subtyping resolution could provide information on classical typing designations, such as serotypes, or be extended beyond current molecular subtyping resolving capabilities to detect single nucleotide differences across entire genomes (69).

1.5.2.1 WGS for *Salmonella* Serotype Prediction

As surveillance and diagnostic laboratories continue to look towards WGS to carry out their mandates, they require the use of applications that can maintain compatibility with their existing systems while also harvesting the full potential of WGS data analysis techniques. With that said multiple applications for the *in silico* determination of *Salmonella* serovars directly from WGS data have been developed in recent years. These applications work in two ways, either by inferring the serotype of an isolate using patterns from other subtyping techniques or by directly determining the antigenic formula of an isolate from the genes present in its genome (76).

1.5.2.1.1 WGS Inference of *Salmonella* Serotypes Using 7-Gene MLST

Previous research has shown that STs and eBGs from the 7-gene MLST scheme for *Salmonella* correlated with many individual serovars, and it has been proposed that STs and eBGs could be used to replace traditional serotyping. It has been shown that for many serovars, such as Typhimurium and Enteritidis, there was a direct correlation between a single eBG and a single serovar; while others serovars, such as Newport, were found to be polyphyletic and instead had representatives within multiple eBGs. Proponents of replacing traditional serotyping

designations with ST information have highlighted its benefits, such as its adequate reflection of the evolutionary groupings of isolates (76). However, traditional serovar designations have a long history of use and familiarity within the larger scientific community so MLST is instead looked upon as a method that could be used to infer traditional serovar designations versus directly replacing them (77). In fact, the WHO-Salm was expected to update the WKL in December 2015 with MLST data for each serovar; however, this new edition of the WKL has been delayed (78). Currently, the MLST database for *Salmonella* is compiled by the University of Warwick and is available via their online database (http://mlst.warwick.ac.uk/mlst), that provides an updated list linking serotypes to the ST generated by 7-gene MLST (76).

While traditional 7-gene MLST protocols did not utilize WGS data, various bioinformatic applications have emerged to mine 7-gene MLST data from WGS data. These approaches include ones that call allele data directly from the raw reads or from *de novo* assembled genomes (79, 80). Raw read analysis uses a read mapping approach to call alleles, where the raw sequencing reads are mapped to a database of alleles. This methodology is utilized by the SRST2 (Short Read Sequencing Typing) program (79). Allele calls can also be made from *de novo* assembled genomes using BLAST (Basic Local Alignment Search Tool) and specialized WGS-based allelic databases for MLST data such as BIGSdb (Bacterial Isolate Genome Sequencing Database) (80). PHE is currently using the 7-gene MLST to infer serovar designations from WGS data generated for *Salmonella* isolates, providing a high throughput, accurate, robust, and reliable typing method that maintains serotyping system in the age of WGS data (77).

1.5.2.1.2 WGS-Based Antigenic Determination Using SeqSero

SeqSero is a web-based application (http://www.denglab.info/SeqSero) developed by researchers from the University of Georgia and represents the other way for determining

Salmonella serotypes from WGS data – specifically determining the antigens directly from the genes carried by the isolate. Specifically, SeqSero looks at the genes encoded by the *rfb* region, for O-antigen determination, and the *fliC* and *fljB* genes, for H1 and H2 antigen determination, respectively. The resulting antigenic formula is looked up in the WKL scheme to provide an overall serovar name. SeqSero can determine the serovar of an isolate either directly from raw sequencing reads or from genome assemblies (42).

When raw sequencing reads are uploaded to the SeqSero website, the reads are mapped to curated databases of the *wzx* and *wzy* genes for O-group determination, and the gene that has the highest number of reads mapped to it determines the specific O-group. For *fliC* and *fljB* an alternative approach is needed due to the high sequence similarity between these genes that can diminish or eliminate the excess number of reads that would be expected to be mapped to the correct gene. For H-antigen determination, raw reads are mapped to the curated H-antigen gene databases in three successive mapping steps that narrow down the gene database to a list of representative alleles for a single antigen cluster (i.e. the g complex of antigens). Following antigen complex determination, the mapped reads from the final mapping are aligned to an individual allele using BLAST, and the gene that has the highest BLAST score provides the antigen prediction (42).

When assembled genomes are uploaded to the SeqSero website the relevant pieces of information from the assembled genome are pulled out and compared to the curated antigen databases using BLAST. To determine the serogroup the *gnd* and *galF* genes, which flank the *rfb* region, are first located in the assembled genome. If both *gnd* and *galF* are found on the same contig the *rfb* region residing in between them is extracted and aligned to the *rfb* database using BLAST, and the serogroup is determined from the highest scoring match. If *gnd* and *galF* are

found to reside on two separate contigs, the two contigs of interest are split at these genes, producing four fragments. These four fragments are aligned to the *rfb* database using BLAST, and once again the highest scoring match from these four fragments determines the serogroup. To determine the H-antigens, SeqSero first performs an *in silico* PCR to pull out the *fliC* and *fljB* genes. The products of this *in silico* PCR are then aligned to the curated H-antigen database using BLAST, where once again the gene determinant with the highest score determines the antigen (42).

1.5.2.1.3 Combined Approach for WGS-Based Serotype Prediction Using SISTR

The *Salmonella in silico* Typing Resource (SISTR) is a web-application that can provide an *in silico* way of determining serotypes and was developed by researchers from PHAC. This application combines both a gene determination and phylogenetic inference methodology to determine an isolate's serotype from a *de novo* assembled draft genome to perform its analysis (41). Like SeqSero (42), SISTR performs an O- and H-antigen prediction through a BLAST search using curated database of alleles. For O-antigen determination SISTR looks at the highly variable *wzx* and *wzy* genes; while the H-antigen genes are once again determined from the *fliC* and *fljB* genes. However, ambiguity may be present in many of the O- and H-antigen predictions (41). For example, many of the alleles from the g-complex of H1 antigens show remarkable sequence similarity (81). As well, the O-antigen prediction is only to the serogroup level and not the individual antigen level, as this is the most precise level possible based on our current knowledge of *Salmonella* O-antigen genetics (82, 83), and sequence similarities between serogroups still pose challenges. Together, this means that ambiguities may be present in any antigen-based prediction from SISTR, as multiple serovars could be possible. To resolve these potential ambiguities SISTR has incorporated a secondary serovar prediction methodology

through phylogenomic inference (41).

Expanding on the 7-gene MLST analysis for *Salmonella* serovar prediction, SISTR incorporates a novel 330 locus core genome MLST (cgMLST) scheme to provide a phylogenetic context to the antigen prediction, resolve any ambiguities present, and act as a confirmation of antigen prediction. Alleles across the 330 loci are called against the SISTR allele database using BLAST. Hierarchical clustering of the cgMLST results, using the pairwise distances between genomes, is used to provide the phylogenetic context for serovar prediction. The predominant serovar within the cgMLST cluster, from which an unknown is placed, is used to provide the cgMLST serovar prediction, and this is used to choose the most likely serovar from the list of potential serovars given from the antigen prediction stage. Together these the two methodologies, antigen prediction and cgMLST serovar prediction, are used to provide an overall SISTR serovar prediction that shows improved concordance with traditional methods than just using an *in silico* antigen prediction method alone. Dendrograms of the cgMLST clustering results can also be visualized and downloaded from the SISTR website (41).

1.5.2.2 WGS for Outbreak Detection and Investigation

The application of WGS data for high-resolution surveillance activities such as outbreak detection and investigation relies on the construction of whole genome phylogenies. The construction of these phylogenies requires the comparative analysis of genomes to produce a phylogenetic tree, which provides a visual representation of how the genomes are related to each other. Much like a real tree, a phylogenetic tree consists of leaves (also called terminal nodes), representing the isolates or sequences being studied, that are connected through branches. Internal nodes are branch points within the tree and are representations of hypothetical ancestors. Highly related isolates or sequences will share internal nodes or recent common ancestors.

Ultimately phylogenetic analysis provides the highest level of resolution possible for determining the relatedness of two or more isolates. Multiple various WGS analysis applications exist for outbreak detection and investigation, but ultimately all applications fall within three general categories, alignment-free, alignment-based, and gene-by-gene methodologies, that all differ in how the comparison between genomes is made (73).

The application of whole genome phylogenies for outbreak detection and investigation is an area of active research and development within the global public health community. Since 2013, PulseNet USA has been prospectively sequencing all *L. monocytogenes* isolates as part of their routine investigations. WGS data is analyzed and compared to the results from PFGE. Work from this initiative has led to an increased number of detected clusters, linked more clusters to food sources, and reduced the median cluster size in comparison to when PFGE results were used alone. Overall this initiative has been highly successful for *L. monocytogenes* and further pilot projects are underway for other priority pathogens (84). The PulseNet International community is also experimenting with the use of WGS to carry out its respective mandates (85) and PulseNet Canada is currently examining two applications for the analysis of WGS data for outbreak detection and investigation. Specifically, they are looking at analysis platforms that use the alignment-based and gene-by-gene methodologies to produce whole genome phylogenies (Personal Communication, Celine Nadon, 2015).

1.5.2.2.1 Single Nucleotide Variant Phylogenomics Pipeline

The Single Nucleotide Variant Phylogenomics (SNVPhyl) pipeline is a bioinformatic pipeline uses an alignment-based approach for molecular epidemiology (86). SNVPhyl and other alignment-based methodologies are computationally intensive processes that generate a multiple sequence alignment (MSA). MSAs align the sequence from each isolate under study into rows

with columns representing a new data point (in this case a nucleotide) of potential homology. Non-homologous data points between the sequences in a MSA represent single nucleotide variants (SNVs), and the greater the number of SNVs separating two sequences the less closely related they likely are (73). The SNVPhyl pipeline exists as a workflow on Galaxy and uses both novel and pre-existing open source bioinformatic tools to generate the MSA, find SNVs, and build the phylogenetic tree (86).

The SNVPhyl pipeline requires the use of a high-quality reference genome. The reference genome may either be a draft or a fully finished genome, but should be of sufficient similarity to the genomes under analysis. After inputting the reference file and the sequence reads of the isolates under study into the pipeline, internal repeats in the reference genome are identified using the NUCmer program provided within the MUMmer (v3.23) package and these regions of the reference are masked from further analysis (86), as they can complicate the read mapping process (73). Next, the sequence reads from the isolates are aligned to the reference genome using SMALT (v.0.7.5), and genomes below a user-defined coverage cutoff are flagged. Two different tools, FreeBayes (v.0.9.20) and SAMtools' BCFtools (version 1.3), are used to call variants across the mapped genome. The variants called from these two methods are consolidated and filtered to ensure the SNVs called are of high quality. The quality of the SNVs are assessed in two ways, ensuring they are not sequencing errors or artifacts. First SNVs are assessed to ensure they were found in an area that met a user-defined minimum mean mapping level, and second SNVs are assessed to ensure the ratio of the reads that called the SNV to those that did not was above a user-defined level. SNV calls that pass these quality metrics are defined as high-quality SNVs (hqSNVs), and these are further scanned to identify and mask recombination events, which are defined by areas with an unusually high density of SNVs (86). As an asexual

process, recombination events do not follow normal hereditary evolution and will greatly alter phylogenies if they are not properly taken into account (73).

The validated hqSNVs are collected and are used to generate a SNV-based MSA that is then run through PhyML (version 3.1), a tree building tool based on maximum-likelihood principles (86). The maximum-likelihood approach for producing phylogenies is a computational intensive character-based tree building methodology that examines the actual variations within the sequence data to infer the best tree possible (73). Phylograms produced from maximum-likelihood approaches are also considered to be highly accurate depictions of molecular phylogenies (87), and provide a wealth of information allowing for the tracing of evolution at each site in the alignment (88).

1.5.2.2.2 Whole Genome Multilocus Sequence Typing

Whole genome MLST (wgMLST) represents an extension of both the 7-gene, and cg-MLST methodologies for molecular epidemiology that incorporates information from the whole genome; thereby, providing a higher level of analysis. In fact, both 7-gene and any cgMLST based systems have limited discriminatory power to perform the necessary phylogenetic analysis needed in outbreak detections and investigations (76, 89). Like other MLST-based systems, wgMLST requires organism-specific schema that defines both the loci and alleles (73, 80). A wgMLST scheme for *Salmonella* has been developed as a plug-in for the BioNumerics suite of software by Applied Maths. This scheme includes over twelve thousand loci that capture diversity both within the core and accessory genome of *Salmonella*. It was developed from 260 *Salmonella* reference genomes, from a variety of common serovars to capture the maximum amount of diversity in the entire genome (90).

In the BioNumerics wgMLST plug-in, alleles are called across the loci using a consensus-based approach, where the same allele call needs to be detected from two methodologies for inclusion as a consensus allele. Calls are first made through an assembly-free approach that uses a k-mer based method to call alleles directly from the raw sequencing reads. The second allele calling method requires a *de novo* assembled genome from which calls are made through a BLAST-based search (http://www.applied-maths.com/applications/wgmlst). Phylogenetic analysis from MLST data is often carried out by determining the pairwise distances between each genome (73). Phylogenetic trees produced through distance methods, such as via Unweighted Pair Group Method with Arithmetic Mean (UPGMA), are much less computationally intensive but provide little information outside of the resulting tree (88).

Sequence reads from online or local servers, such as IRIDA, are linked to the BioNumerics database allowing for a fast and efficient software system, and any metadata uploaded to the software remains within the local environment to ensure confidentiality. All of the analysis, from the *de novo* assembly through to the allele calls, occurs within the BioNumerics software environment, through the BioNumerics external calculation engine. This calculation engine can be either a subscription-based cloud service or a locally installed computer cluster, allowing for a light and efficient system that does not require massive local hardware infrastructure (http://www.applied-maths.com/applications/wgmlst).

1.6 Hypothesis

The hypotheses for this research are two-fold and relate to how WGS data can be used and incorporated into the current surveillance system for *S. enterica* in Canada. I hypothesize that *Salmonella* serovars can be predicted from WGS data with accuracy and this test is suitable for implementation in a reference laboratory. As well, I hypothesize that subtyping by WGS

improves accuracy and confidence in the categorization of cases in an outbreak scenario, including a multi-strain event.

1.7 Objectives

The objectives of my research were multifaceted. To address hypothesis one, first I set out to uncover any gaps in the *in silico* serotype prediction tool SISTR, for important serovars within the Canadian clinical context to enhance and improve upon the tool development. Next, I worked to validate the multiple *in silico S. enterica* serotyping tools in comparison to traditional serotyping techniques. To address hypothesis two, I set out to compare the ability of PFGE and WGS data analysis tools for the identification and investigation of a multi-strain *S. enterica* outbreak associated with sprouted chia powder. Lastly, I aimed to assess the impact of WGS data analysis tools on Canada's current surveillance system for *Salmonella* and to develop considerations that can guide future implementation of this technology.

Chapter 2: Materials and Methods

2.1 WGS-Based Prediction of *Salmonella* Serovars

2.1.1 Isolate Selection

Three panels of isolates were selected for this study, each with their own criteria. In total, all three panels contained 813 isolates from 142 ssp I serovars, 8 ssp II serovars, 8 ssp IIIa serovars, 11 ssp IIIb serovars, 8 ssp IV serovars, and 2 *Salmonella bongori* serovars. Together these isolates were selected to provide coverage of the important clinical serovars in Canada and provide coverage for the majority of antigenic determinants currently described. Use of Canadian strains allowed for a confirmatory check of any *in silico* results with traditional serotyping, something that could not be completed using public data sets.

2.1.1.1 Panel One Isolate Selection

Panel one consisted of 400 isolates from twenty serovars (Supplementary Table A) and was selected to provide a cross-section of isolates most frequently encountered in the Canadian clinical landscape, as together these serovars represent about 85% of all reported cases of human salmonellosis in Canada (27). For an isolate to be considered for this panel it must have satisfied the following inclusion criteria: (i) submitted to the National Microbiology Laboratory (NML) Enterics division between the years 2009-2013, and (ii) been among the top twenty clinical serovars in Canada[1](27). Isolates were excluded from this list if they were designated as a reference or proficiency test isolate. From the generated list, twenty isolates from each serovar were randomly selected using a random number generator (https://www.random.org) after the total population of isolates had been filtered by serotype and sorted by reported isolate number.

[1] Top twenty clinical serovars in Canada include: Enteritidis; Heidelberg; Typhimurium; ssp I 4,[5],12:i:-; Thompson; Infantis; Newport; Typhi; ssp I 4,[5],12:b:-; Brenderup; Saintpaul; Javiana; Paratyphi A; Hadar; Agona; Paratyphi B variant Java; Stanley; Oranienburg; Muenchen; and Montevideo (27).

2.1.1.2 Panel Two Isolate Selection

Panel two consisted of 92 isolates representing a collection of clinically relevant but infrequently encountered serovars (Supplementary Table A). The 46 serovars represented in this panel included serovars: with an increased association with extra-intestinal[2] or travel related infections[3]; those deemed difficult to differentiate by traditional serotyping[4]; those from the clinically relevant non-subspecies I subspecies[5]; and those left untypeable by traditional serotyping[6]. For an isolate to be considered for this panel it must have satisfied the following inclusion criteria: (i) submitted to the NML enterics division between the years 2009-2013, and (ii) been among the clinically relevant but infrequently encountered serovars. Isolates were excluded from this list if they were designated as reference or proficiency test isolate. From the generated list, a designated number of isolates from each serovar were randomly selected using a random number generator (https://www.random.org) after the total population of isolates had been filtered by serotype and sorted by reported isolate number.

2.1.1.3 Panel Three Isolate Selection

Panel three[7] consisted of 321 isolates chosen to represent the most globally prevalent *Salmonella* serovars from both human and non-human sources as well as to provide coverage for

[2] Five serovars with increased association of extra intestinal infections were chosen, specifically Cerro, Dublin, Panama, Sandiego, and Schwarzengrund (27).
[3] One serovar with an increased association of travel related infections was chosen, specifically Corvallis (27).
[4] Six serovars were chosen as they were deemed difficult to differentiate by traditional techniques, specifically Paratyphi B, Senftenberg, Kouka, Carrau, Madelia, and Lattenkamp (Unpublished data, NML Reference Services, 2015).
[5] Clinically relevant non-subspecies I serovars came from subspecies, II, IIIa, IIIb, and IV (Unpublished data, NML Reference Services, 2015).
[6] Isolates left untypeable by traditional serotyping included representatives from subspecies I, II, IIIa, IIIb, and IV that were either completely untypeable or had a rough-O designation.
[7] Work conducted on panel three from isolate selection up to data analysis was carried out by collaborators at NML Guelph.

the majority of antigenic determinants currently described. The isolates that were included in this panel were previously used in validation studies on other molecular based serotyping methods (91). These comprised an additional 152 isolates from the serovars (and or subspecies) targeted in panels one and two, plus an additional 167 isolates from 97 serovars not represented in panels one or two, and 2 isolates from *S. bongori*, (Supplementary Table A) all of which later were referred to as the non-target serovars.

2.1.2 Genome Sequencing and Assembly

2.1.2.1 DNA Extraction

For isolates from panels one and two, genomic DNA was extracted using either the Qiagen DNeasy 96 Blood and Tissue Kit (Qiagen, Mississauga, ON, Canada) from overnight cultures grown in Luria Bertani (LB) Lennox 0.5% NaCl broth, or using the Epicenter Metagenomics DNA isolation kit for water (Mandel Scientific Company, Guelph, ON, Canada) from isolated colonies grown on a nutrient agar plate. For both kits, manufacturer's instructions were followed. For isolates from panel three, genomic DNA was extracted from isolated colonies on overnight LB agar plates using the KingFisher Cell and Tissue DNA Kit (VWR, Mississauga, ON, Canada) on the KingFisher Flex (VWR), or using the EZ1 DNA tissue kit and BioRobot (Qiagen). Again, manufacturer's instructions were followed with the addition of lysozyme (Sigma-Aldrich Canada, Oakville, ON, Canada) to a concentration of 10mg/ml in the cell lysis incubation stage.

2.1.2.2 Library Construction and Genomic Sequencing

Recovered DNA from all isolates was quantified using the Qubit DNA quantification system (Invitrogen Canada, Burlington, ON, Canada) and diluted to a genomic DNA concentration of 0.2 ng/µl. Library construction was completed using the MiSeq Nextera XT

library preparation kit (Illumina, San Diego, CA, USA), and the libraries were size selected for a minimum insert size of 500 bp using the BluePippin (Sage Science, Beverly, MA, USA). The Illumina MiSeq with the MiSeq Reagent Kit v3 600 cycle (2x300 bp forward and reverse) was used to perform the paired-end sequencing of all isolates. Sequencing data for all isolates was performed by the Genomics Core (NML) and was uploaded and stored on the IRIDA platform in a single project dedicated to the prediction of *Salmonella* serovars using *in silico* tools.

2.1.2.3 WGS Quality Control

Sequence reads for each isolate were run through the FastQC Summary Report workflow on Galaxy (version 1.0) (Bioinformatics Core, NML). This workflow utilizes the FastQC:ReadQC tool (version 0.10.1) (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/), which provides a simple analysis of the raw sequencing data for each isolate. Outputs from this tool are summarized and tabulated by the workflow to provide an easy to interpret summary report. Isolates from panels one and two with an estimated genome coverage greater than 30X (assessed using a total sequence length of 5 Mbp), proceeded to the next stage of analysis, isolates that failed to reach these standards were sent back for resequencing. Isolates from panel three with an estimated genome coverage greater than 15X proceeded to the next stage of the analysis. The average estimated coverage for all our isolates was 89X.

Isolates that reported with an estimated genome coverage greater than 200X were downsampled, so as not to overload subsequent assembly tools. Downsampling was completed using the seqtk_sample tool in Galaxy (version 1.0-r75-dirty) (Heng Li, Broad Institute), which will reduce the number of reads through random subsample generation to provide a desired coverage level. For our analysis, we set our desired coverage threshold for these isolates to 100X, with five isolates requiring downsampling.

2.1.2.4 WGS Assembly

The *de novo* assembly of the isolates from panels one and two was carried out using the SPAdes assemblies with FLASH workflow on Galaxy (version 1.2) (Bioinformatics Core, NML). This workflow utilizes both FLASH (Fast Length Adjustment of SHort reads) (version 1.2.9), which produces longer merged reads for a less computationally intensive and an overall better assembly (92) and SPAdes (St. Petersburg genome Assembler) (version 3.5.0), which utilizes paired de Bruijn graph method to assemble the genomes (93). The careful correction tool option was selected in SPAdes with k values set at 21,33,55,77,99,127. The workflow also filters out any short, low coverage, or repeated contigs that were assembled by SPAdes. The filtering parameters were set to remove contigs that fell below 1Kb length cutoff, below one-third of the average coverage, and were 75% above the average coverage.

To assess the quality of the assemblies, statistics from the workflow were downloaded and any isolate with over 150 contigs or with less than a 4 Mbp assembled genome was flagged for reanalysis. Reanalysis involved running the contigs from the assembly through blastn on the National Center for Biotechnology Information (NCBI) website (https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE_TYPE=BlastSearch) to check for contamination, isolates with hits to other organisms were re-sequenced. The assembled genomes of isolates from panels one and two had on average 42 contigs (range 13 to 147) and produced genomes with an average size of 4,728,551 bp (range 4,404,162 to 5,759,342).

The paired-end reads of isolates from panel three were first merged using FLASH (version 1.2.9) (92), and then *de novo* assembled into contigs using SPAdes (version 3.5.0) (93). SPAdes was run using the careful option to correct assembly errors. The assembled genomes of isolates from panel three had on average 83 contigs (range 27 to 949) and produced genomes

with an average size of 4,770,399 bp (range 4,323,398 to 6,101,829).

2.1.3 Initial SISTR Validation and Gap Finding

All isolates from panels one and two were analyzed using SISTR (41) and results were collated for the purposes of uncovering gaps within the system, including technical, coding, or biological errors, and to develop recommendations for the refinement of the program. Multiple bioinformatic tools were utilized to explore the results and pinpoint these areas. All findings from this phase were communicated to collaborators responsible for the SISTR program, who made the necessary corrections to the SISTR algorithms or source code.

In Galaxy, the NCBI BLAST+ tool kit (version 2.2.29) (94, 95) was used extensively to explore the allelic calls that were initially returned from SISTR. The SISTR allelic databases were downloaded in December 2015 from the SISTR Bitbucket repository (https://bitbucket.org/peterk87/sistr_backend) and were converted to nucleotide BLAST databases using the makeblastdb Galaxy tool (version 0.1.00). At times reference genomes on NCBI; previously published PCR-based primers and their products; or other allelic databases such as the SeqSero allelic databases, downloaded from the SeqSero website (http://www.denglab.info/SeqSero), served as the basis for custom BLAST databases. Assembled isolates of interest were queried against the BLAST databases using the megablast search as part of the blastn Galaxy tool (version 0.1.00).

The SMALT toolkit (version 0.7.5) (http://www.sanger.ac.uk/science/tools/smalt-0) and SAMtools (version 0.1.19) (96) were used to map raw reads of the isolates to reference sequences. The smalt index tool on Galaxy (version 1.0.0) (http://www.sanger.ac.uk/science/tools/smalt-0) was used to prepare smalt indexes of reference sequences for read mapping investigations. Raw sequencing reads from isolates of interest were

mapped against the smalt index using the smalt map tool on Galaxy (version 1.0.0) (http://www.sanger.ac.uk/science/tools/smalt-0)(96). Default settings on the smalt map tool was used. User defined parameters were set as follows: the maximum and minimum insert size was set to 1000 and 20 bp, respectively; the identity threshold for a mapping to be reported was set to 0.5; and if multiple mappings with the same alignment score was reported, one was picked at random. The resulting pileups were displayed in Tablet (version 1.15.09.01) (97).

R (version 3.0.2) (https://www.r-project.org), bedtools (version 2.18.2) (98), and SAMtools (version 0.1.18) (96) were used as part of the panel coverage report tool on Galaxy (version 0.0.2) to produce a more detailed report on the raw read coverage for select isolates. Previously developed read maps produced using the smalt map tool were converted to a BAM file format using the SAM-to-BAM tool on Galaxy (version 1.1.4) (96) and the resulting BAM files were input into the panel coverage report tool, along with a BED file of the reference. The BED file of the reference sequence was produced via the Fasta to Bed File tool on Galaxy (version 1.0.0) (Bioinformatics Core, NML).

At times dendrograms produced from the pairwise cgMLST similarity matrix between our genomes and publicly available genomes within the SISTR database were also visualized and downloaded from the SISTR website for examination and analysis.

2.1.4 Comparison of *in silico* Serotype Prediction Methods

Following the refinement of the SISTR platform, serovar predictions for all isolates from panels one, two, and three were generated. Isolates were uploaded to both the SISTR (https://lfz.corefacility.ca/sistr-app/) (41) and SeqSero (http://www.denglab.info/SeqSero)(42) websites for serovar prediction from these platforms. To generate the 7-gene MLST serovar prediction, the 7-gene ST generated through the SISTR platform (41) was compared to the

University of Warwick *Salmonella enterica* MLST database (http://mlst.warwick.ac.uk/mlst/) (downloaded December 2015) to specifically link serovars to ST generated (76).

All methods were compared to traditional serotyping results using the interpretation criteria described below. *In silico* serotyping results were categorized as either full, partial, genotypic, or incorrect matches in comparison to traditional serotyping. Full matches occurred when the serovar prediction was concordant with the reported serovar by traditional typing. Partial matches were identified as serovar predictions that required further information to get a full serovar designation as either the prediction was ambiguous (multiple serovars were indicated) or was missing information. Genotypic matches occurred when the genomic serovar prediction was incompatible with the reported phenotypic serovar, specifically due to the carriage of antigenic genes that were not phenotypically expressed. Lastly, matches were considered incorrect when the overall serovar prediction was not concordant with the reported serovar by traditional serotyping, after the traditional serotype was confirmed a second time. All full, partial, and genotypic matches were considered successful results, as these categories were indicative of a positive result in relation to traditional serotyping.

2.1.4.1 Statistical Analysis of Platforms

Differences in the number of successful results from each platform were evaluated for statistical significance using a Fisher's exact test. A 2-by-2 contingency table was created via GraphPad Quickcalcs (www.graphpad.com/quickcalcs). A P value of less than or equal to 0.05 was considered significant.

For serovars Enteritidis and Typhimurium, the test sensitivity and specificity for the *in silico* serovar prediction from each platform in comparison to traditional serotyping was also assessed through a 2-by-2 table analysis (99). In determining the test sensitivity and specificity

all genotypic matches were first removed from the analysis due to the incongruence between the phenotypically and genetically derived test results. The inclusion of these results would artificially reduce the measured sensitivity and specificity of the methods, due to the inherent incompatibility of the genomic and phenotypic methods for these isolates. Test sensitivity was defined as TP/(TP+FN) and test specificity was defined as TN/(TN+FP), where TP = true positives, FN = false negative, TN = true negatives, and FP = false positives.

2.2 Examination of WGS for *Salmonella* Outbreak Investigation

2.2.1 Selection of Outbreak

All closed PulseNet Canada *Salmonella* clusters from 2010 to 2016 were examined for clusters in which (i) an Outbreak Investigations Coordinating Committee was activated under FIORP, and (ii) multiple serovars were linked to the outbreak. Clusters were excluded from this list if they contained representatives from serovars Enteritidis, Heidelberg, or Typhimurium, as these serovars were part of other WGS retrospective investigations at the time (Personal Communication, Celine Nadon, 2016). A single outbreak from 2014 linked to sprouted chia powder in Canada met the above criteria and was selected for further examination by WGS.

2.2.1.1 Isolate Selection

The sprouted chia powder outbreak consisted of a total of 63 cases of *Salmonella* from four serovars (Hartford, Newport, Saintpaul, and Oranienburg). The first human isolate identified as part of the outbreak was uploaded to the PulseNet Canada National *Salmonella* PFGE database in BioNumerics (version 6.01) (Applied Maths, Belgium) on January 30th, 2014 and the outbreak was officially declared over on July 29th, 2014 (http://www.phac-aspc.gc.ca/phn-asp/2014/salmonella-nh-053114-eng.php). To capture the full outbreak and the relevant non-outbreak case comparators, all isolates from these four serovars uploaded to the PulseNet Canada

PFGE database during this time period, as well as in the 60 days before and after, were selected for sequencing. PFGE pattern upload date was the only consistently reported date and was therefore used to determine the temporally related comparators. A total of 371 isolates from these four serovars were identified and selected for examination by WGS.

2.2.2 Genome Sequencing

2.2.2.1 DNA Extraction

Of the 371 isolates initially selected for sequencing, only 368 isolates were actually sequenced. Four isolates were removed at this stage in the analysis as the stocks could not be located, the stock was non-viable, or the sample that was pulled from the stock was found to be not *Salmonella*. DNA was extracted using either the Qiagen DNeasy 96 Blood and Tissue Kit (Qiagen) from overnight cultures grown in LB-Lennox 0.5% NaCl broth, or using the Epicenter Metagenomics DNA isolation kit for water (Mandel Scientific Company) from isolated colonies grown on a nutrient agar plate. For both kits, the manufacturers' instructions were followed.

2.2.2.2 Library Construction and Sequencing

Recovered DNA from all isolates was quantified using the Qubit DNA quantification system (Invitrogen Canada) and diluted to a genomic DNA concentration of 0.2 ng/µl. Library construction was completed using the MiSeq Nextera XT library preparation kit (Illumina). The Illumina MiSeq with the MiSeq Reagent Kit v3 600 cycle (2x300 bp forward and reverse) was used to perform paired-end sequencing of all isolates. Sequencing of all isolates was performed by the Genomics Core (NML) and WGS data was uploaded and stored on the IRIDA platform in a single project dedicated to the retrospective analysis of this outbreak.

2.2.2.3 WGS Quality Control

Sequence reads for the 368 isolates that were sequenced, were run through the FastQC

Summary Report workflow on Galaxy (version 1.0) (Bioinformatics Core, NML) (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/), as previously described. Isolates with an estimated genome coverage greater than 28X (assessed using a total sequence length of 5 Mb) and with a most abundant read length of 300 bp, proceeded to the next stage of analysis, isolates that failed to reach these standards were sent back for resequencing. The average estimated genome coverage for all isolates was 76X. All isolates were then *de novo* assembled using the SPAdes assemblies with FLASH workflow on Galaxy (version 1.2) (Bioinformatics Core, NML)(92, 93) as previously described and uploaded to SISTR (41) to confirm the serovar designation. The assembled genomes contained on average 33 contigs (range 14 to 182), and were on average 4,737,084 bp in length (range 4,520,788 to 5,308,824) and serotypes for all isolates were confirmed through SISTR.

2.2.3 WGS-Based Data Analysis

2.2.3.1 SNVPhyl Analysis

The SNVPhyl Paired-End Galaxy workflow (version 1.0) (Bioinformatics Core, NML) (86) was run to generate the various core genome SNV phylogenies generated in this study. Reference genomes consisted of closely related (same serovar) closed genomes found in the NCBI repositories, or a *de novo* SPAdes assembled genomes of high quality that were produced in the quality control step if no closed genome was readily available. The minimum coverage during read mapping for each isolate was set to 10X across 80% of the reference genome to ensure high mapping quality. Minimum mean mapping was set to 30X coverage to ensure that the base on which a variant call is made had sufficient coverage. Lastly, the alternative allele proportion was set to 0.75 to ensure that the ratio of reads that contained the SNV to those that did not was above 0.75.

The Rearrange SNV Matrix tool on Galaxy (version 1.7.0) (Bioinformatics Core, NML)(86) was then run using the newick tree, the pseduoalignment Phylip file, and the SNV matrix generated from the SNVPhyl pipeline as inputs. This tool converts the branch lengths in the SNV tree from the number of substitutions per site to the total amount of SNVs. The outputted newick file along with a metadata file was then uploaded to the Microreact phylogenomic visualization webpage (https://microreact.org/upload) to visualize the trees.

2.2.3.2 wgMLST Analysis

Raw sequencing data files for all isolates were linked from the IRIDA project to the BioNumerics software program (version 7.6.1) (Applied Maths). The wgMLST profiles for all isolates were calculated using the wgMLST plug-in and run using the BioNumerics external calculation engine. The assembly-free allele calls were made using a k-mer size of 35 and the minimum coverage at any allele was set to 3X. The contigs used in the assembly-based allele calls were selected to ensure they had a minimum length of 1000 bp. Following calculation of the wgMLST experimental data, allelic comparisons of isolates were created for relevant groups and dendrograms were generated from the pairwise distances between isolates within a comparison via UPGMA hierarchical clustering method.

2.2.4 Case Categorization and Cluster Identification from WGS Data

Any clades containing outbreak cases were identified in the trees and defined as an outbreak cluster. The minimum and maximum SNV or allele differences between the isolates within these outbreak clusters were identified and additional non-outbreak isolates that grouped within these outbreak clusters were considered outbreak related by WGS analysis. Additional WGS-based clusters were also identified within the trees and defined as any group of two or more isolates that were separated by less than the maximum SNV or allele differences seen in the

outbreak clusters. Other PFGE-based clusters identified among the study isolates were also examined to see if they formed WGS-based clusters.

Chapter 3: Results

3.1 WGS-Based Prediction of *Salmonella* Serovars

3.1.1 Initial SISTR Validation and Gap Finding

All sequences from panels one and two were initially analyzed by SISTR to uncover any gaps in the *in silico* serotype prediction and to develop recommendations for the refinement of the program. Direct matches between the SISTR serovar prediction and the serovar identified by traditional methods were reported for 69.7% of the 492 isolates tested. These results were investigated in-depth to develop a series of recommendations and fixes.

3.1.1.1 O-Antigen Troubleshooting

A total of 65 isolates returned without an O-antigen prediction, resulting in no serovar prediction for these isolates. For all 65 isolates, H-antigen and cgMLST prediction were congruent with the phenotypic serovar prediction. The lack of O-antigen prediction was explored further, and two issues were found to explain these results. First, a technical error was uncovered in the computer code for the prediction of the C2-C3 serogroup. The technical error was uncovered when a custom BLAST database, consisting of the *wzx* and *wzy* genes that SISTR uses, was created and returned with a positive hit to the assembled genomes for thirteen C2-C3 serogroup isolates. This indicated that the genes SISTR uses for an O-antigen prediction were present in the assemblies for these isolates, and a technical issue with SISTR was preventing the call. As a result, a misplaced period in the code was corrected.

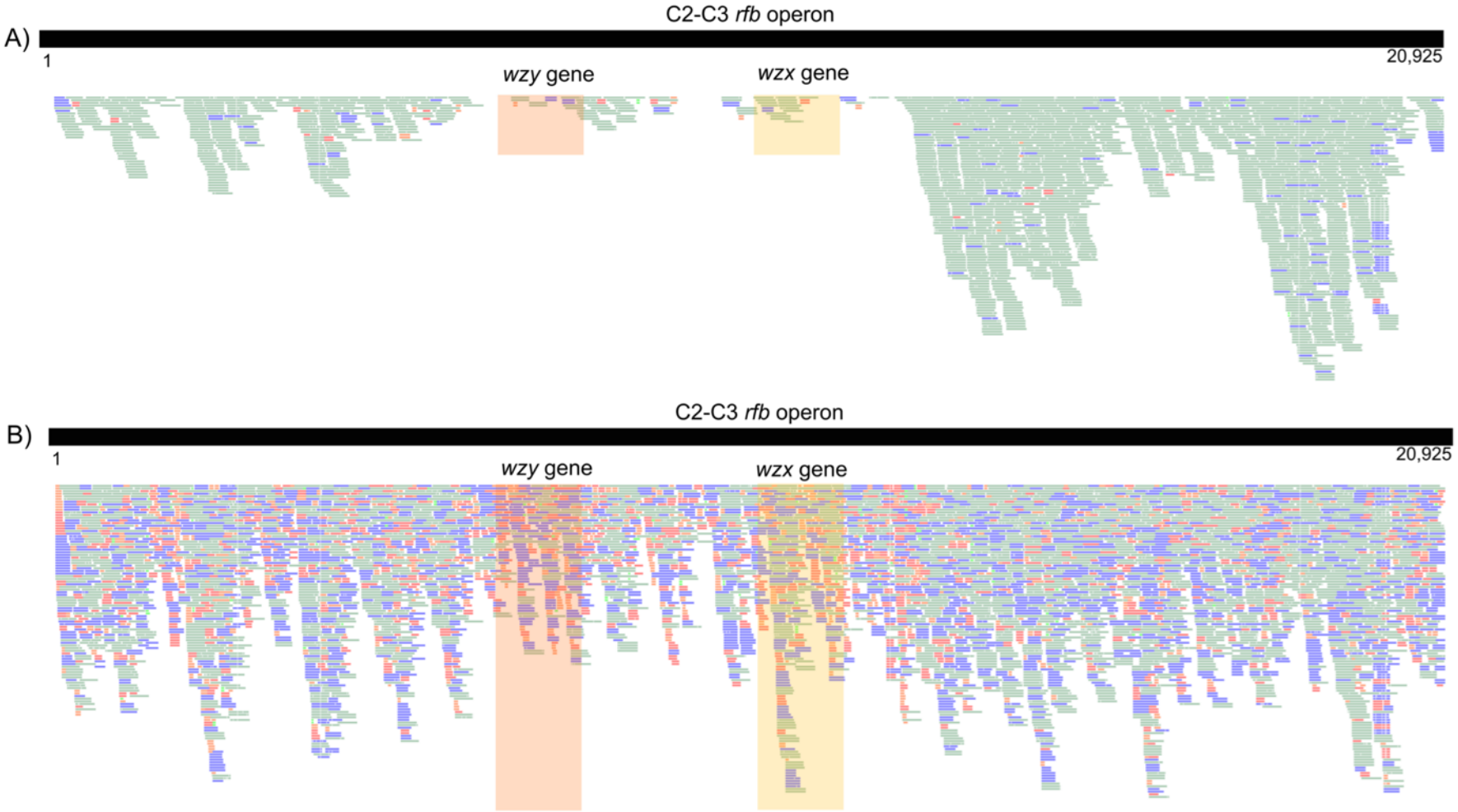
The second issue with O-antigen prediction was uncovered which lead to a lack of O-antigen prediction in 52 isolates. For these isolates, the *wzx* and *wzy* genes that SISTR uses for an O-antigen prediction were missing from the assemblies, and the larger *rfb* region in which these genes lie was split over two contigs. Raw sequencing reads for these isolates were read-mapped to reference *rfb* regions for their respective serogroups. Reference *rfb* regions for the individual

serogroups were pulled from the SeqSero *rfb* cluster database, which was downloaded in December 2015. All pile-ups were found to display the same pattern, which showed very low (<5X) to no sequencing read coverage over large parts of the *rfb* region, including at the *wzx* and *wzy* genes (Figure 1A). To assess whether the low sequencing coverage was responsible for the lack of *in silico* O-antigen prediction, three isolates were resequenced using the same procedures as described in the methods, but the prepared genomic libraries were not size selected for a minimum insert size of 500bp on the BluePippin. These newly generated WGS data produced pile-ups over the *rfb* region with a much higher and more uniform coverage, including at the *wzx* and *wzy* genes (Figure 1B). Panel coverage reports across the *rfb* region for these isolates were also generated using data that were both size-selected and not size-selected. The size-selected data showed: a lower average base coverage, a greater number of non-covered bases, and a lower percentage of bases under 30X coverage, over the *rfb* region in comparison to the non-size-selected data (Table 2). This indicates the poor suitability of size-selected data for the *in silico* determination of *Salmonella* serotypes.

Table 2: Results from the panel coverage report for the *rfb* region of three *S. enterica* isolates using data that was generated with and without a library size selection step for a minimum insert size of 500 bp.

Serovar	Serogroup	Length of *rfb* region	Size-Selected	Average Base Coverage	Number of Non-Covered Bases (bp)	% of Bases at 30X Coverage
Enteritidis	D1	19394	Yes	21X	1535	0
			No	51X	0	88.33
Newport	C2-C3	20925	Yes	24X	1680	0
			No	61X	0	94.26
Thompson	C1	9388	Yes	27X	3134	0
			No	48X	0	60.34
Average			**Yes**	**24X**	**2116**	**0**
			No	**53X**	**0**	**80.98**

Figure 1: Raw read pileup for a *Salmonella* Newport isolate mapped to a C2-C3 *rfb* reference operon and visualized using Tablet. **(A)** Pileup for this isolate when the sequencing library underwent a size selection step prior to data generation. **(B)** Pileup for the same isolate when the sequencing library did not undergo a size selection step. Read colour corresponds to read length: red ≤150; blue ≥150 but ≤250; green ≥250. The *wzy* and *wzx* genes are highlighted in orange.


A)
C2-C3 rfb operon
1
20,925
wzy gene
wzx gene
B)
C2-C3 rfb operon
1
20,925
wzy gene
wzx gene

3.1.1.2 Recognizing Incompatibilities Between Phenotypic and Genomic Data

The initial validation of SISTR immediately brought to light the issue of incompatible phenotypic vs genotypic results for a select group of isolates. The incompatibilities resulted from the carriage of unexpressed antigenic determinants. In total 32 of the 492 isolates tested displayed incompatible results between their phenotypic and genetic serotyping. Twenty-five of these isolates had a Rough-O or untypeable designation by traditional serotyping yet provided a serovar call through SISTR; while, seven were traditionally serotyped as a monophasic serovar, only to be found to genetically carry the *fljB* gene.

For the Rough-O and untypeable serovars, no single genetic change or phylogenetic signal was detected among the 25 rough isolates tested in the study to distinguish them from isolates that express antigens. For the seven isolates that carried the *fljB* gene but did not express it, multiple mutations in the flagellar phase variation machinery were detected. Previously characterized (100, 101) loss of expression mutations in the flagellar phase variation machinery were converted to BLAST databases which were queried against the seven isolates. Evidence of these previously defined loss of expression mutations were found in all seven of the monophasic isolates that genetically carried the *fljB* gene (Table 3). Together these incompatible isolates highlight that *in silico* and traditional serotyping assess different elements, which can provide different results that are both considered correct.

Table 3: Previously characterized mutations in the flagellar phase variation machinery that lead to the loss of *fljB* expression for seven traditionally serotyped monophasic isolates identified during the SISTR validation.

Reported Serovar	*In silico* Serovar	Mutation found	Number of isolates	Source
4,[5],12:i:-	Typhimurium	Truncated *hin* Possible *IS*26 element	1	Boland *et al*. (100)
4,[5],12:b:-	Paratyphi B	Single base pair deletion at position 2,915,827	1	Toboldt *et al*. (101)
4,[5],12:b:-	Paratyphi B	Complete loss of *hin* gene	5	Toboldt *et al*. (101)

3.1.1.3 cgMLST clustering

A total of 50 out of the 492 isolates tested in the initial SISTR validation displayed poor cgMLST clustering results. Poor cgMLST clustering occurred when the closest genome to which these isolates clustered matched to less than 85% of the 330 alleles in the cgMLST scheme. While the majority of these isolates were from the rare and unusual serovars tested, some were from common serovars, such as serovar Oranienburg. Poor cgMLST clustering indicates the need to expand the cgMLST database to fully encompass these unique lineages within Canada.

3.1.1.4 Paratyphi B, Paratyphi B var. Java and its Monophasic Variant

Initial SISTR serovar prediction was unable to differentiate the two serovar variants of antigenic formula 4,[5],12:b:1,2, denoted Paratyphi B and Paratyphi B variant Java (13, 102), as all isolates with this antigenic formula returned with a Paratyphi B serovar prediction. The variants of this antigenic formula are traditionally differentiated by their ability to ferment *d*-tartrate. Paratyphi B isolates cannot ferment *d*-tartrate and are considered to be dT-, while Paratyphi B var. Java isolates can and are dT+(13, 102). This distinction is considered to be clinically important as dT- isolates are linked to a more severe typhoid-like disease, while dT+ isolates are linked to a less severe gastrointestinal disease. PCR-based detection of *d*-tartrate fermentation status was previously developed by targeting of the *STM3356* gene. Isolates that

were dT+ display an ATG start codon at the start of the *STM3356* gene, while those that are dT- display an ATA codon in this spot, resulting in a non-transcribed and null-function *STM3356* gene (102).

A 273bp probe sequence, taken from a Paratyphi B var. Java isolate from this study, was designed from the 5' end of primer 167 from Malorny *et al* (102) and ending with the ATG start codon of the *STM3356* gene. This probe sequence was used to create a custom BLAST database. Isolates with a reported serovar of Paratyphi B or Paratyphi B var. Java were queried against this database to assess their *d*-tartrate fermentation status. Isolates that returned with a BLAST result that aligned over the entire 273bp were labeled dT+, those that aligned over 272bp were labeled dT-, as these isolates displayed an ATA codon at the end of the sequence. All isolates with a reported serovar of Paratyphi B returned with a dT- status via this *in silico* test, while all isolates with a reported serovar of Paratyphi B var. Java returned with a dT+ status, confirming the correct result.

A 330 loci cgMLST dendrogram was produced from all isolates (n=40) in the SISTR database with the reported serovar of Paratyphi B (n=28) and Paratyphi B var. Java (n=12) (Figure 2A). No clear phylogenomic signal is apparent between these isolates based on their reported serovar status. A 330 loci cgMLST dendrogram produced from all isolates sequenced from panels one and two with the reported serovars of Paratyphi B (n=5) or Paratyphi B var. Java (n=20) was produced (Figure 2B). In this tree, all Paratyphi B isolates form a tight cluster and are separated from the Paratyphi B var. Java isolates.

The sequences for the 40 isolates in the SISTR database were downloaded from NCBI in April 2014 and assembled to determine their *in silico d*-tartrate fermentation status. Only 7 of the 28 isolates with a reported serovar of Paratyphi B were found to actually be dT- using this *in*

silico assessment, while all isolates with the reported serovar of Paratyphi B var. Java were dT+. The *in silico* assessment of *d*-tartrate fermentation status indicates that a large number of the public genomes were incorrectly called Paratyphi B. A SISTR generated 330 loci cgMLST cladogram was produced for all 85 isolates with the genetically determined antigenic formula of B:b:1,2 (Figure 2C). This dendrogram incorporates the *in silico d*-tartrate fermentation status; thereby showing the true capabilities of the SISTR cgMLST schema for differentiating serovar variants Paratyphi B and Paratyphi B var. Java. A clear phylogenetic signal of genotypically dT- or Paratyphi B isolates can be seen in this dendrogram.

The serovar 4,[5],12:b:-, most often thought of as a monophasic variant of Paratyphi B var. Java, was not initially in SISTR, and all genotypically compatible isolates (n=14) with this serovar designation were considered to be serovar Schleissheim or inconclusively labeled as Schleissheim| II 4,[5],12,[27]:b:[e,n,x]. Literature searches revealed that Schleissheim is an incredibly rare serovar with the antigenic formula 4,[5],12,27:b:- and only three human cases of disease due to this serovar have been reported (103-105).

Two genomes in the cgMLST database had the reported serovar of Schleissheim; however, investigations into these genomes revealed 7-gene MLST ST that were associated with the monophasic Paratyphi B var. Java isolates, and not serovar Schleissheim (76). As well, an additional eleven other isolates with reported serovars of Paratyphi B var. Java (n=6), Paratyphi B (n=4), and 4,[5],12:b:- (n=1) were curated in the SISTR database to serovar Schleissheim, following an antigen prediction of B:b:-. Once again 7-gene MLST ST revealed that these isolates were not serovar Schleissheim, but were monophasic Paratyphi B var. Java, indicating that errors within the SISTR database were conflating serovar Schleissheim with monophasic Paratyphi B var. Java isolates.

Figure 2: Dendrograms of *S. enterica* subsp. *enterica* isolates with the antigenic formula 4,[5],12:b:1,2 produced from the 330 loci cgMLST scheme in SISTR. **(A)** Dendrogram of the 40 public genomes with a reported serovar of Paratyphi B (Orange) or Paratyphi B var. Java (Purple). **(B)** Dendrogram of all isolates from panels one and two with a reported serovar of Paratyphi B (Red) or Paratyphi B var. Java (Blue). **(C)** Dendrogram of all 85 isolates with the genetically determined antigenic formula of B:b:1,2, nodes are coloured coded by the initial reported serovar and source, and boxes are overlaid to show *in silico d*-tartrate fermentation status.

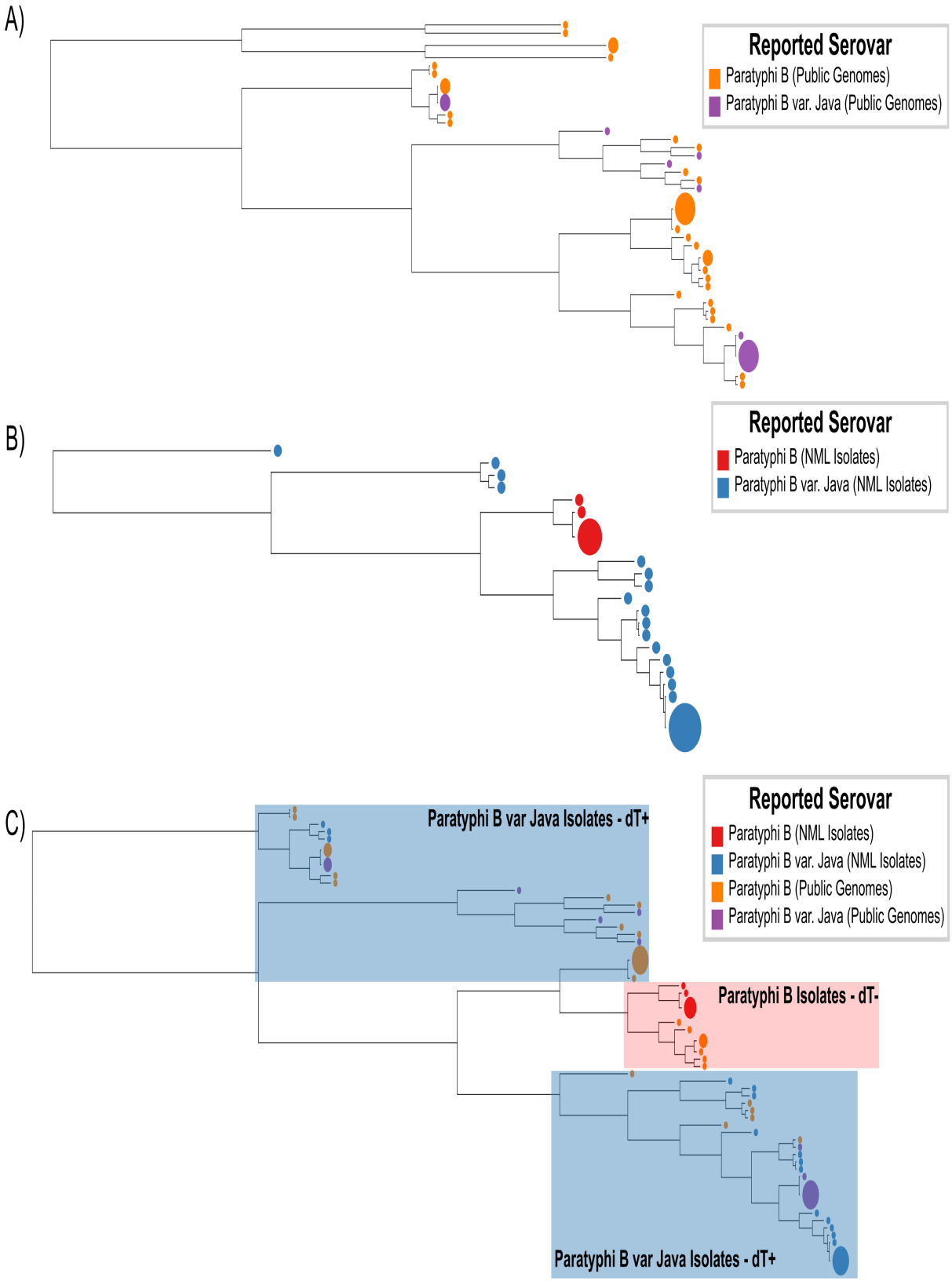
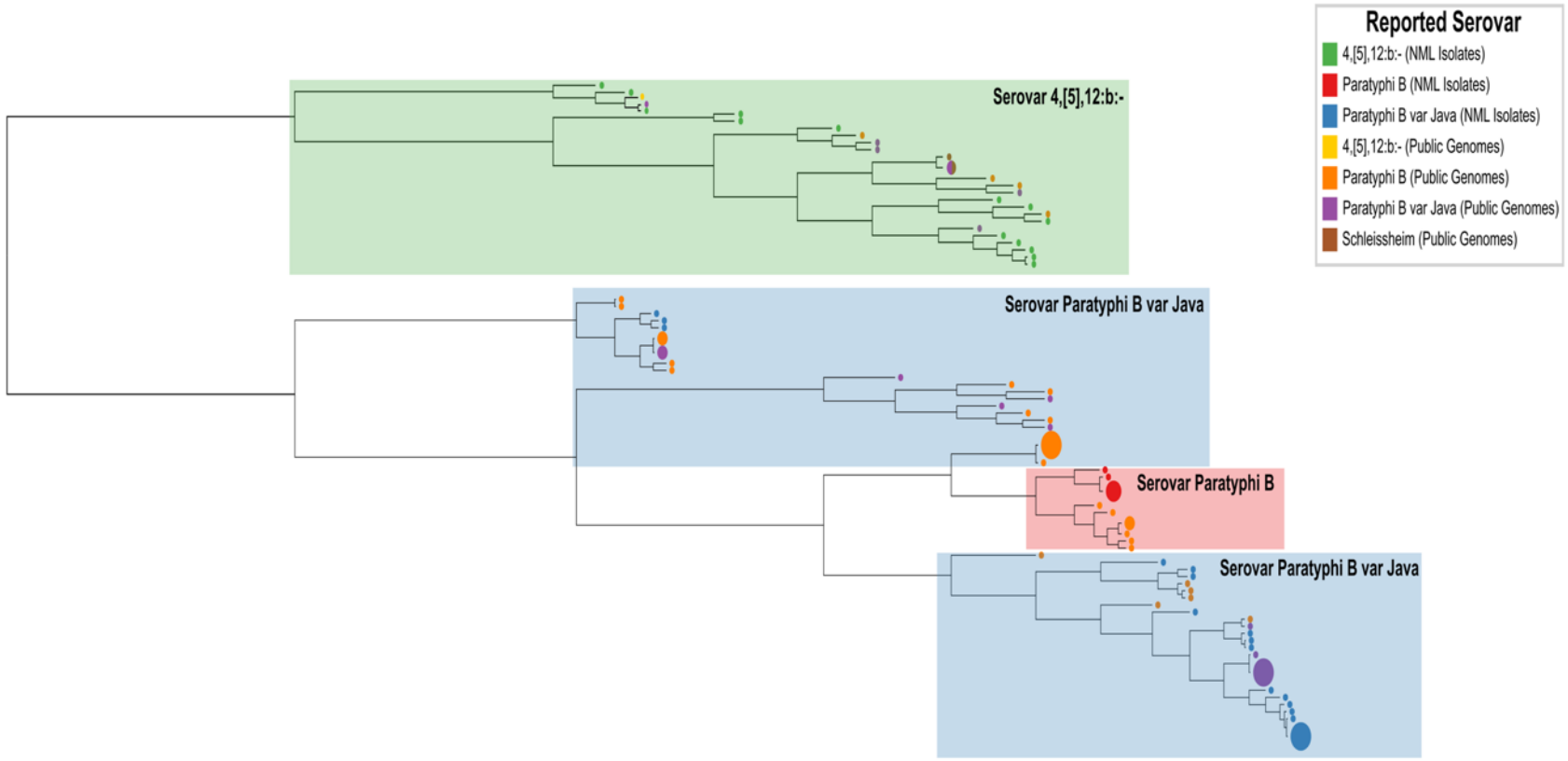

A)
Reported Serovar
Paratyphi B (Public Genomes)
Paratyphi B var. Java (Public Genomes)
B)
Reported Serovar
Paratyphi B (NML Isolates)
Paratyphi B var. Java (NML Isolates)
C)
Paratyphi B var Java Isolates - dT+
Reported Serovar
Paratyphi B (NML Isolates)
Paratyphi B var. Java (NML Isolates)
Paratyphi B (Public Genomes)
Paratyphi B var. Java (Public Genomes)
Paratyphi B Isolates - dT-
Paratyphi B var Java Isolates - dT+

Figure 3: Dendrogram of all isolates with the genetically determined antigenic formula of B:b:1,2 and B:b:-, produced from the 330 loci cgMLST scheme in SISTR. Nodes are coloured by the initial reported serovar and isolate source. Colourized boxes are overlaid to show the proposed serovar designations of the lineages.


Reported Serovar
4,[5],12:b:- (NML Isolates)
Paratyphi B (NML Isolates)
Paratyphi B var Java (NML Isolates)
4,[5],12:b:- (Public Genomes)
Paratyphi B (Public Genomes)
Paratyphi B var Java (Public Genomes)
Schleissheim (Public Genomes)
Serovar 4,[5],12:b:-
Serovar Paratyphi B var Java
Serovar Paratyphi B
Serovar Paratyphi B var Java

A SISTR-generated 330 loci cgMLST dendrogram was produced for all isolates from panels one and two as well as any public genomes with the genetically determined antigenic formulas of B:b:1,2 and B:b:- (Figure 3). Clear lineages of 4,[5],12:b:-, Paratyphi B var. Java, and Paratyphi B isolates are apparent.

3.1.1.5 The Non-Subspecies *enterica* Serovars

Initial SISTR results did not differentiate between the two species of *Salmonella* and the five subspecies of *Salmonella enterica*. The lack of differentiation lead to incorrect or inconclusive results for the majority of isolates tested from non-subspecies I serovars. The SISTR generated 330 loci cgMLST dendrogram of all 492 isolates tested as part of panels one and two plus any public sequences of non-subspecies I isolates was produced (Figure 4) and shows that the 330 loci cgMLST scheme is able to properly differentiate the various subspecies of *Salmonella*

3.1.1.6 Flagellar Antigen Issues

Multiple instances of novel genes variants for antigens were found. Specifically, novel gene variants were found for the H1 antigens: c; l,v; l,w; and z35, and the H2 antigen z53. All isolates with novel gene variants resulted in wrong H-antigen calls. The incorrect H-antigen that was called for these isolates displayed several mismatches and/or indels, often with little coverage over the entire gene (Table 4). Phenotypic confirmation of the antigens was performed, and five novel gene sequences for rare flagellar antigens was added to the SISTR flagellar antigen database.

In addition, the *fliC* gene for l,v in the SISTR database was found to actually encode the l,z13,z28 antigen. Two ssp II isolates with the serovar of 58:l,z13,z28:z6 returned with an exact match to the *fliC* gene for the antigen l,v. However, phenotypic confirmation maintained the

l,z13,z28 antigen, and the flagellar antigen database was curated as this gene was relabelled to represent the antigen l,z13,z28.

Table 4: The *fliC* or *fljB* gene matches of six Salmonella isolates that returned with incorrect flagellar antigen prediction from SISTR.

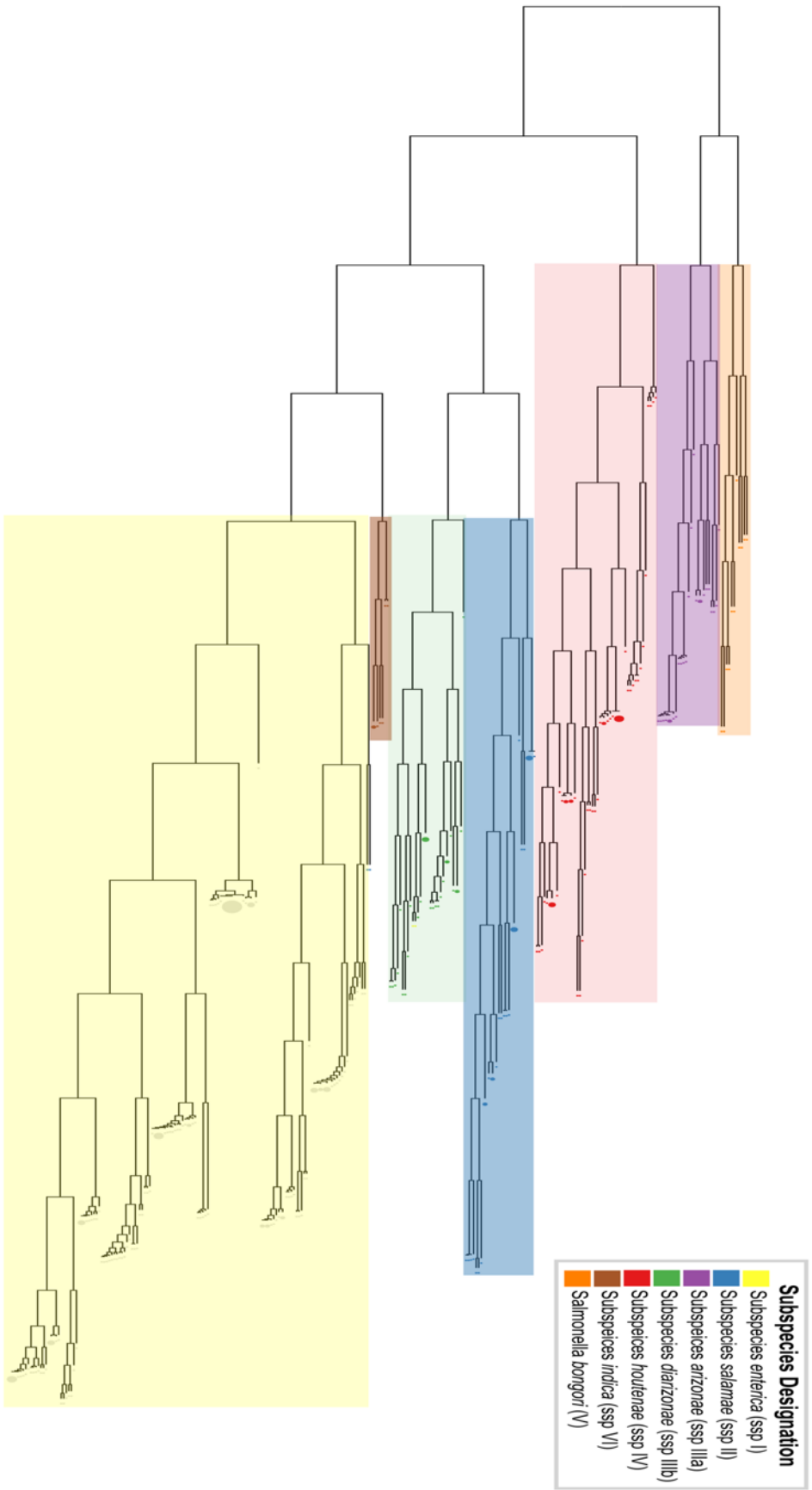
Confirmed Serovar	Number of Isolates	Gene	Antigen Called	% Coverage (bp match)	Number of Mismatches	Number of Gaps[1]
ORough:c:-	1	*fliC*	b	45% (813)	13	0
ssp IIIb 50:l,v:z35	1	*fliC*	z47	100% (1504)	0	4
ssp II D3:l,w:e,n,x	1	*fliC*	b	45% (812)	16	0
Lattenkamp (45:z35:1,5)	2	*fliC*	B	45% (813)	17	1
ssp IIIb 11:k:z53	1	*fljB*	z50	100% (1497)	5	0

[1] Gaps in alignment would have indicated indels within the gene variant and non-triplet indels would have resulted in a frameshift mutation

3.1.2 Comparison of *in silico* Serotype Prediction Methods

A total of 813 isolates were assessed using the three methods for *in silico Salmonella* serovar prediction. SISTR, SeqSero, and 7-gene MLST provided successful results for 94.8%, 88.2%, and 88.3% of the 813 isolates tested, respectively (Table 5). Successful results were considered to be any case in which the prediction did not include incorrect information. These successful results could be further broken down into full, inconclusive, or incompatible matches. Full matches were considered as identical serovar matches between traditional serotyping and the *in silico* method under study, and were found in 89.7%, 54.1% and 77.9% of the isolates tested using SISTR, SeqSero, and 7-gene MLST, respectively. Partial matches were recorded when the information provided by the *in silico* test was insufficient or would require further laboratory testing to narrow down the result, and were found in 1.1%, 30.0%, and 6.4% of the isolates tested using SISTR, SeqSero, and 7-gene MLST, respectively. When the phenotypic and genetic

Figure 4: The SISTR-generated 330 loci cgMLST dendrogram of all 492 isolates from panels one and two plus publically available non-subspecies I genomes. Colourized boxes are overlaid over the various species and subspecies lineages of *Salmonella.*


Subspecies Designation
Subspecies enterica (ssp I)
Subspecies salamae (ssp II)
Subspecies arizonae (ssp IIIa)
Subspecies diarizonae (ssp IIIb)
Subspecies houtenae (ssp IV)
Subspecies indica (ssp VI)
Salmonella bongori (V)

serovar predictions were incompatible due to the carriage of unexpressed antigenic determinants, it was considered a genotypic match; this was found in 4.1% of all isolates tested, regardless of the serovar prediction software used. Lastly, incorrect matches were found in 5.1%, 11.8% and 11.7% of isolates tested using SISTR, SeqSero, and 7-gene MLST, respectively. The number of successful results reported by SISTR was deemed to be significantly greater than the number of successful results reported by either SeqSero or MLST (one-tailed p-values of less than 0.001). However, the observed differences in the number of successful results between SeqSero and MLST was not deemed to be statistically significant. The results from each platform could also be broken down by the serovar analyzed (Supplemetary Table B), with the non-target serovars responsible for the most number of incorrect results from each of the platforms tested.

Table 5: Performance of the three *in silico* methods for *Salmonella* serovar prediction, SISTR, SeqSero, 7-gene MLST, compared to traditional serotyping for 813 *Salmonella* isolates.

Platform	Successful results			Total Successful Results	Incorrect Results
	Full Match[1]	Partial Match[2]	Genotypic Match[3]		
SISTR	729 (89.7%)	9 (1.1%)	33 (4.1%)	771 (94.8%)	42 (5.1%)
SeqSero	440 (54.1%)	224 (30.0%)	33 (4.1%)	717 (88.2%)	96 (11.8%)
MLST	633 (77.9%)	52 (6.4%)	33 (4.1%)	718 (88.3%)	95 (11.7%)

[1] Complete match between *in silico* serovar designation and phenotypic serovar designation
[2] *in silico* serovar designation would require further testing to provide a full result
[3] *in silico* serovar designation is incompatible with phenotypic serovar designation due to carriage of unexpressed antigens.

A total of 26 isolates were designated with a Rough-O antigen by traditional serotyping. These isolates were considered genotypic matches; however, complete serovar calls were generated for 96%, 54%, and 85% of the rough isolates tested by SISTR, SeqSero, and 7-gene MLST, respectively, and partial results were generated for all isolates tested by SISTR, and SeqSero. Predictions across the various platforms were consistent with each other, as were any

reported H antigens. However, one 7-gene MLST prediction was inconsistent with the reported H antigens and the predictions from SISTR and SeqSero.

Sensitivity and specificity were also calculated for each *in silico* method for serovars Enteritidis and Typhimurium, due to their increased global importance (Table 6). SISTR, SeqSero, and 7-gene MLST displayed a sensitivity of 95.2%, 81.0%, and 95.2%, and a specificity of 99.7%, 99.8%, and 99.2%, respectively, for serovar Enteritidis. For serovar Typhimurium, the sensitivities were 100%, 97.4%, and 100%, and specificities were 99.6%, 100%, and 97.2% respectively for SISTR, SeqSero, and 7-gene MLST analysis.

Table 6: Sensitivities and specificities for the prediction of serovars Enteritidis and Typhimurium using three *in silico* methods for *Salmonella* serovar prediction, SISTR, SeqSero, and 7-gene MLST in comparison to traditional serotyping.

	Method	TP[1]	TN[2]	FP[3]	FN[4]	Total Tested	Sensitivity	Specificity
***Salmonella* serovar Enteritidis**	SISTR	40	736	2	2	780	95.2	99.7
	SeqSero	34	737	1	8	780	81.0	99.8
	MLST	40	732	6	2	780	95.2	99.2
***Salmonella* serovar Typhimurium**	SISTR	39	738	3	0	780	100	99.6
	SeqSero	38	741	0	1	780	97.4	100
	MLST	39	720	21	0	780	100	97.2

[1]True positives [2]True negatives [3]False positives [4]False negatives

3.2 Outbreak Investigation

3.2.1 Description of the Sprouted Chia Outbreak

In May of 2014, a national outbreak investigation was initiated for a multi-serovar outbreak of *Salmonella* that was ultimately linked to sprouted chia powder products (106). During the investigation, a case definition, which outlines the specific set of criteria for inclusion in the outbreak (38), was created. This case definition was changed multiple times to reflect the additional serovar and PFGE patterns that were identified in food samples or were deemed similar to the outbreak strains. The final confirmed case definition encompassed four serovars

and thirteen different PFGE patterns (Figure 5) (107); it included two-enzyme matches for previously seen or common PFGE patterns, and one-enzyme matches for unique PFGE patterns never before seen in Canada (Unpublished internal report, PHAC, 2014). In total, 63 confirmed cases were identified between December 4th, 2013 and June 22nd, 2014 (Unpublished internal report, PHAC, 2014)(106).

The identification of the contaminated sprouted chia powder and the production facility source by the United States Centers for Disease Control helped inform the epidemiological, food safety, and laboratory investigation in Canada, as well as the actions taken to end the outbreak. A recall of products containing sprouted chia powder in Canada was initiated on May 30th, 2014 and was expanded an additional eight times throughout the month of June to incorporate additional products from the same processing facility. Multiple recalls were also initiated in the United States (107).

A total of 28 food isolates were collected during the investigation in Canada that had matching PFGE patterns to those included in the official case definition. Additional isolates with unique PFGE patterns with no clinical matches were also found in sprouted chia products (Figure 6). These patterns were not included in the case definition as they were not linked to any human illness (Unpublished internal report, PHAC, 2014).

3.2.2 Overview of WGS Analysis for the Chia Outbreak

Phylogenetic trees, using all 368 isolates that were sequenced as part of this study, were produced in order to evaluate the effectiveness of WGS data analysis techniques in connecting isolates from multiple serovars that were associated with a single food item. The SNVPhyl tree was prepared using the *Salmonella* Newport closed genome, SL256 (downloaded December 2016), as a reference, and was built using 68,473 hqSNVs across 80.13% of the reference

(Figure 7). The minimum spanning tree was also produced using the wgMLST data based on the allelic differences across 9,075 loci (Figure 8). These phylogenetic trees are largely uninformative themselves as they don't provide the in-depth view and level of phylogenetic resolution needed to properly assess the outbreak. However, these trees do show that isolates from the same serovar are more closely related to each other than to isolates from other serovars, and all four serovars examined in this study are to some degree polyphyletic, but have at least one dominant clade that contains the majority of isolates responsible for human disease in Canada. The chia outbreak isolates, denoted in yellow with the outbreak code PNC-Multi-1, reside within their respective serovar groupings. From these trees, it is also apparent that WGS data analysis methods are no better then PFGE in connecting the diverse strains that were found in the sprouted chia products.

3.2.2.1 Serovar Hartford

A total of 46 of the 368 isolates sequenced were from serovar Hartford. Phylogenetic trees were produced, and a single isolate was removed from further analysis of the serovar Hartford isolates due to the large genetic distance (thousands of hqSNVs or alleles) between it and the rest of the group. A SNVPhyl tree was built for comparison using 3,161 hqSNVs across 92.04% of the selected reference genome, the SPAdes assembled isolate Hart-029 (Figure 9). Meanwhile, the wgMLST tree for these isolates was built using allelic character data across 5,039 loci (Figure 10). Twenty-six of these isolates were identified as part of the PNC-Multi-1 outbreak investigation. WGS analysis by both SNVPhyl and wgMLST grouped all 26 of these isolates into a tight cluster that was separated by a maximum of six hqSNVs or alleles (Table 7). One additional isolate grouped with this cluster by WGS analysis. This additional isolate had a matching PFGE pattern to the outbreak patterns, but was excluded from the official outbreak

Figure 5: Confirmed case definition for the multi-jurisdictional outbreak of multiple *Salmonella* serovars linked to sprouted chia products from 2014.

[1] PFGE pattern identified in sprouted chia food samples collected by CFIA/provincial laboratories.
[2] PFGE pattern included through standard PulseNet Canada protocol for similarity.
[3] PFGE pattern included based on identification of food/environmental sample from the United States outbreak.

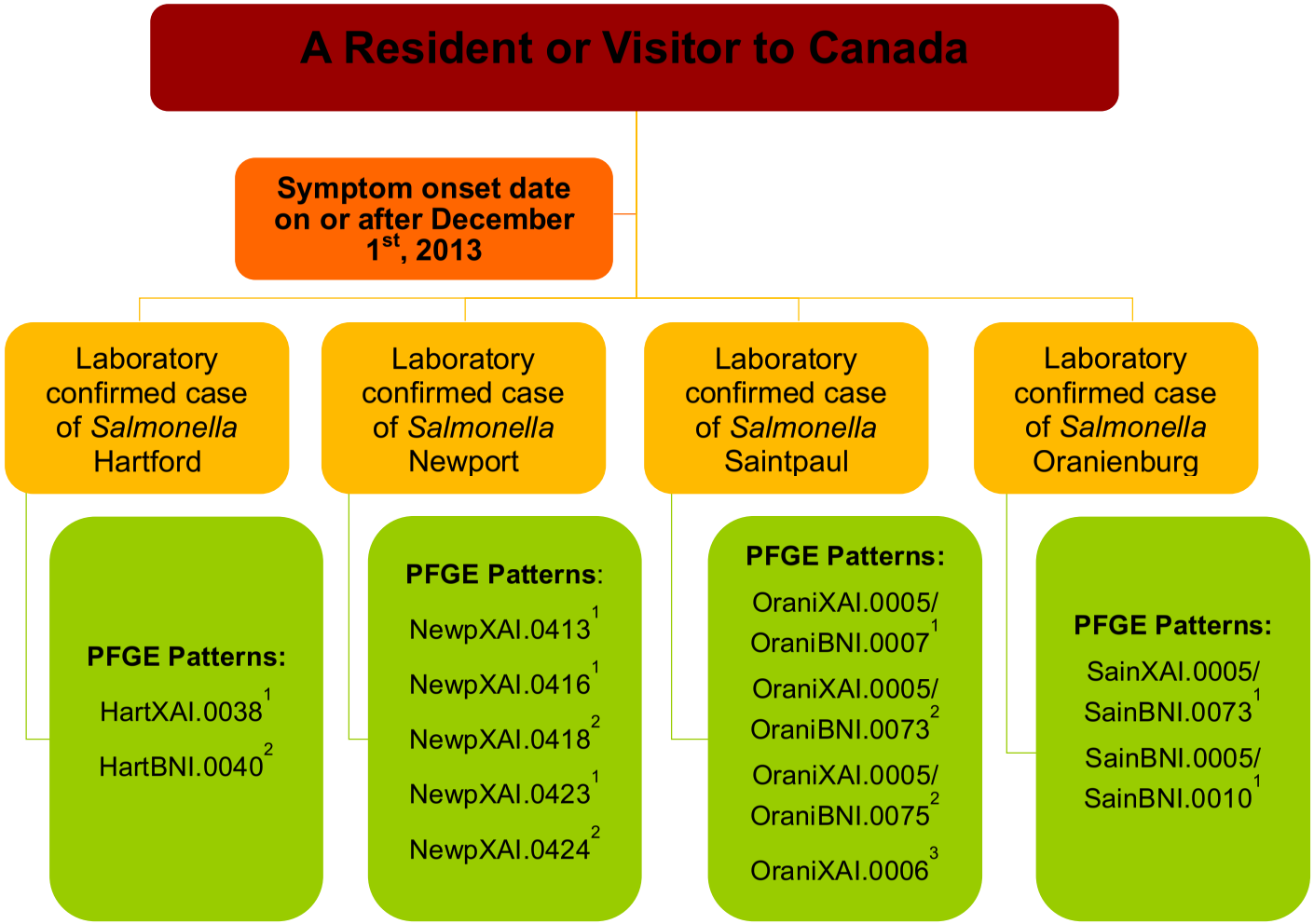
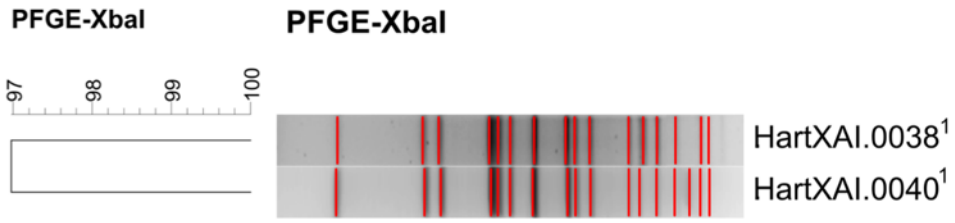

A Resident or Visitor to Canada
Symptom onset date on or after December 1st, 2013
Laboratory confirmed case of Salmonella Hartford
Laboratory confirmed case of Salmonella Newport
Laboratory confirmed case of Salmonella Saintpaul
Laboratory confirmed case of Salmonella Oranienburg
PFGE Patterns:
HartXAI.0038 1
HartBNI.0040 2
PFGE Patterns:
NewpXAI.0413 1
NewpXAI.0416 1
NewpXAI.0418 2
NewpXAI.0423 1
NewpXAI.0424 2
PFGE Patterns:
OraniXAI.0005/ OraniBNI.0007 1
OraniXAI.0005/ OraniBNI.0073 2
OraniXAI.0005/ OraniBNI.0075 2
OraniXAI.0006 3
PFGE Patterns:
SainXAI.0005/ SainBNI.0073 1
SainBNI.0005/ SainBNI.0010 1

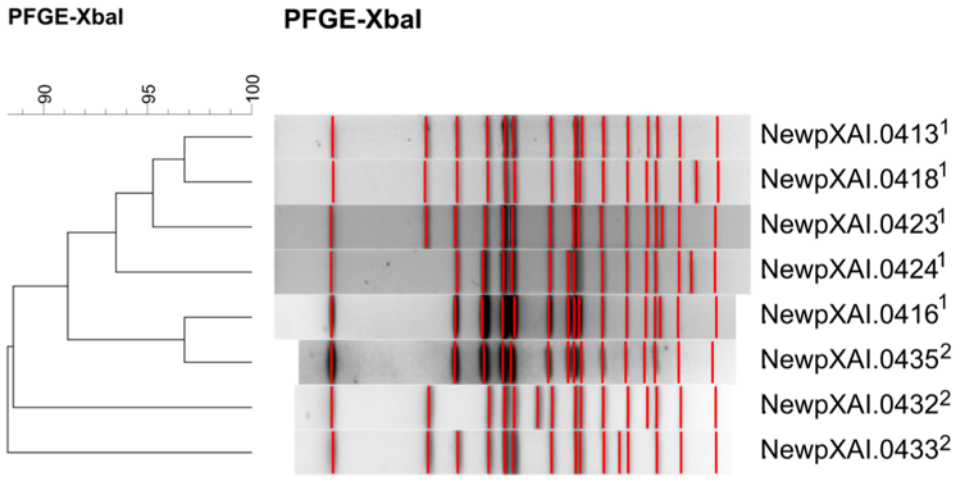
Figure 6: *Xba*I PFGE patterns from the *Salmonella* outbreak associated with sprouted chia products. Scale indicates the percent similarity determined on the basis of Dice coefficients, and cluster analysis was performed by UPGMA. Red lines were overlaid over the specific bands in the pattern for better visualization.

[1]Pattern identified in the case definition.
[2] Pattern identified in sprouted chia food samples collected by CFIA/provincial laboratories, but not linked to any human clinical disease.

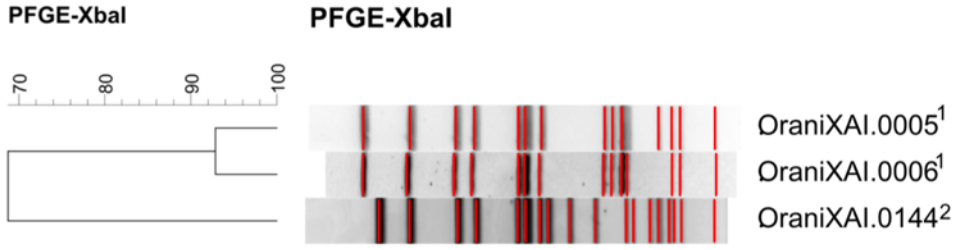
Salmonella Hartford



Salmonella Newport



Salmonella Oraienburg



Salmonella Saintpaul

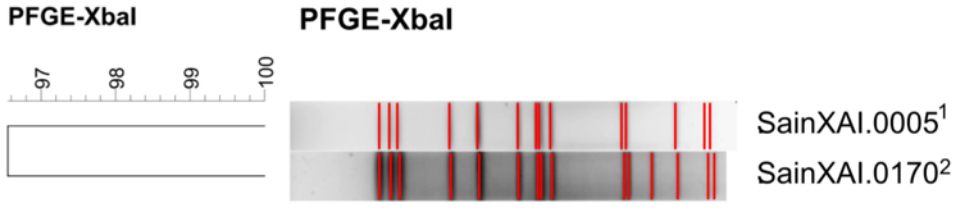


Figure 7: SNVPhyl tree of 368 *Salmonella* isolates sequenced from the outbreak period. Tree was built using 68,473 hqSNVs across 80.13% of the reference genome, the closed *Salmonella* Newport isolate SL256. Each node is colour coded by serovar with isolates from the chia outbreak appearing yellow, regardless of serovar. Shapes denote isolate source (food or clinical). Scale bar indicates the number of hqSNVs.

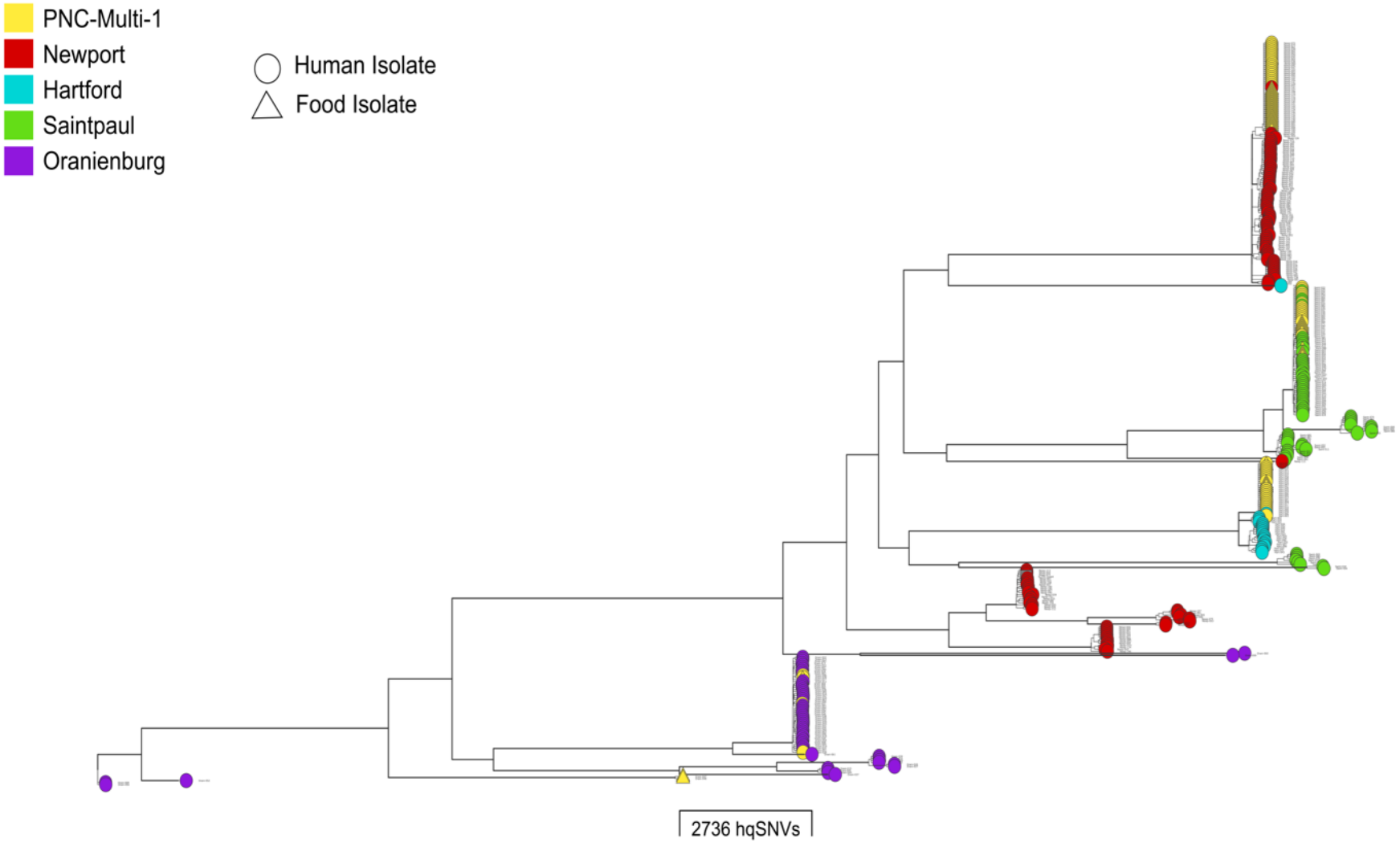

PNC-Multi-1
Newport
Hartford
Saintpaul
Oranienburg
Human Isolate
Food Isolate
2736 hqSNVs

Figure 8: wgMLST analysis, represented as a minimum spanning tree, of the 368 *Salmonella* isolates sequenced from the outbreak period. Tree was built using character data on 9,075 loci. Each node is colour coded by the isolates serovar designation. Isolates from the chia outbreak are displayed in yellow regardless of serovar.

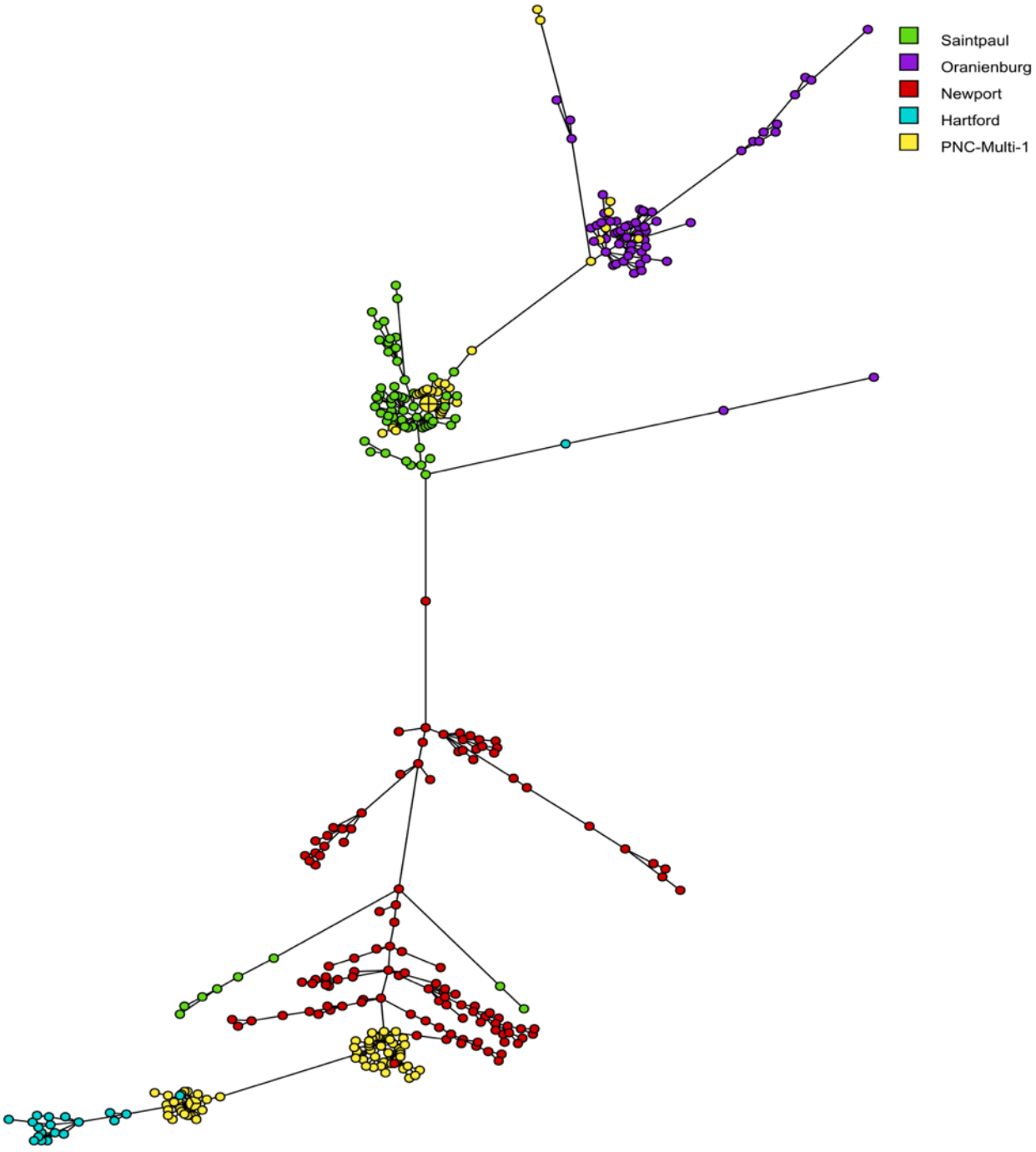

Saintpaul
Oranienburg
Newport
Hartford
PNC-Multi-1

Figure 9: SNVPhyl tree of 45 *Salmonella* Hartford isolates from the time period surrounding the chia outbreak. Tree was built using 3,161 hqSNVs across 92.04% of the reference genome, the SPAdes assembled isolate Hart-029. Each node is colour coded by PFGE-based clusters with non-cluster isolates appearing blue. Node shape corresponds to the source of the isolate. WGS-based clusters are boxed with SNV differences defining the cluster beside the cluster. Scale bar indicates the number of hqSNVs.

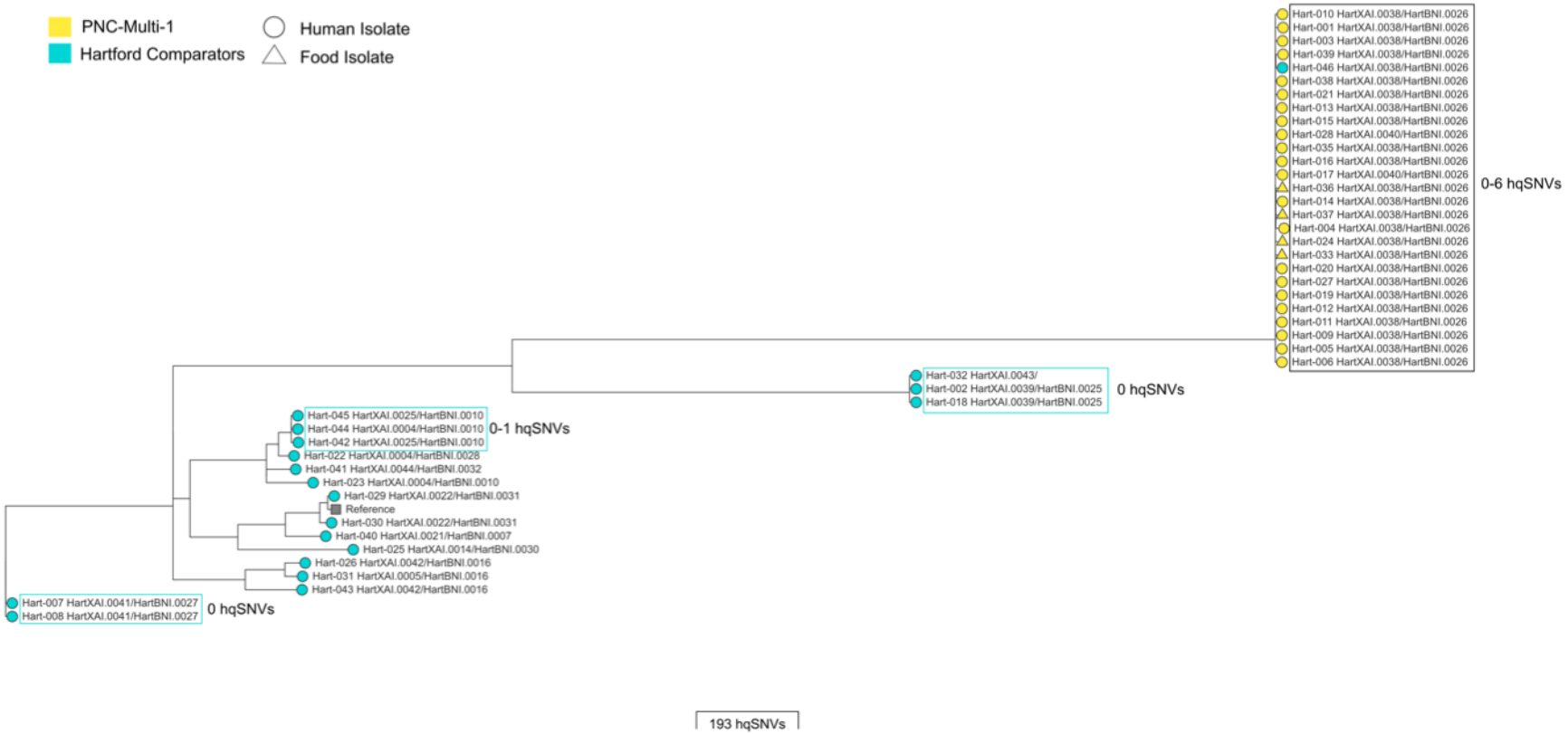
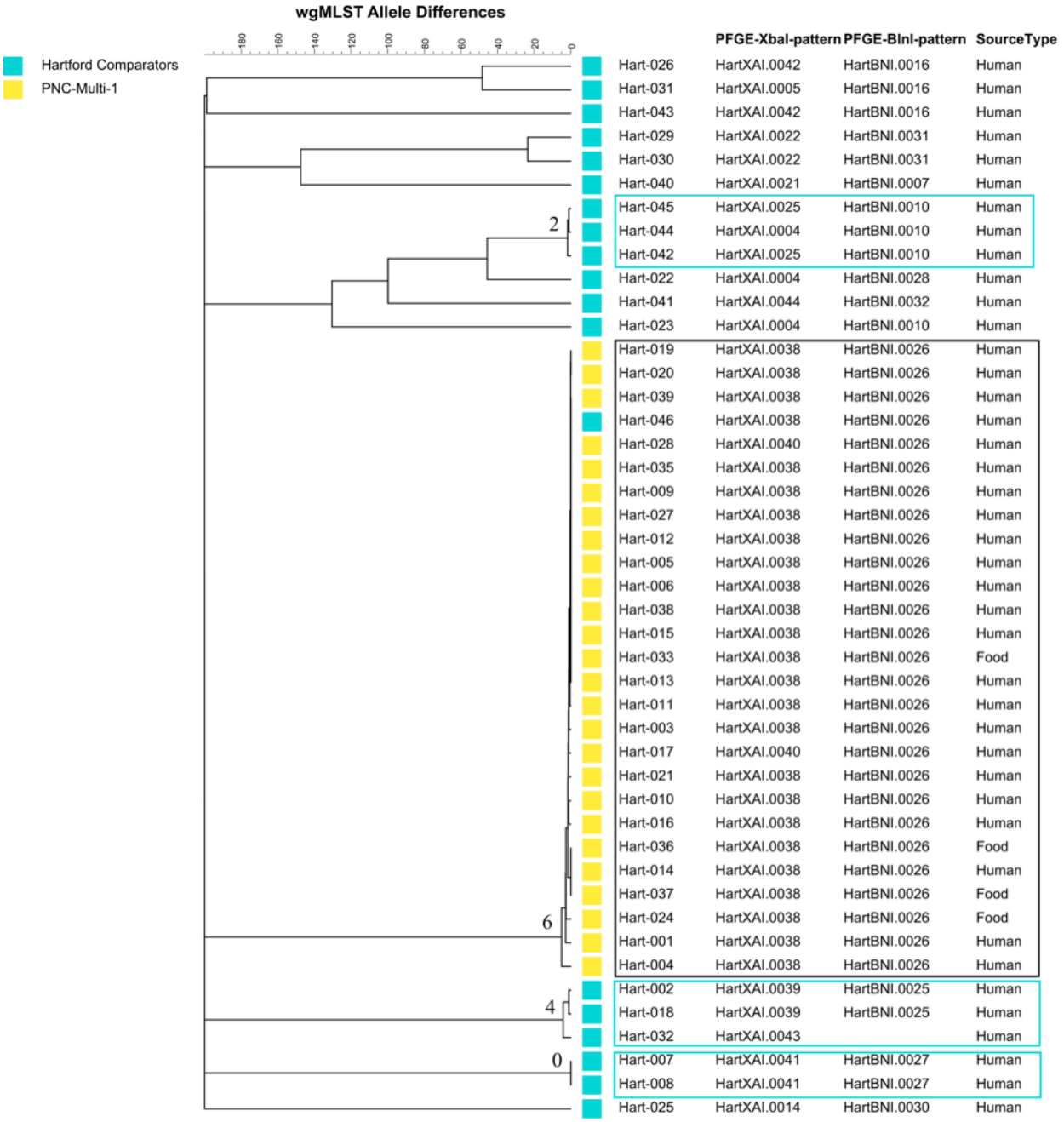

PNC-Multi-1
Hartford Comparators
Human Isolate
Food Isolate
Hart-010 HartXAI.0038/HartBNI.0026
Hart-001 HartXAI.0038/HartBNI.0026
Hart-003 HartXAI.0038/HartBNI.0026
Hart-039 HartXAI.0038/HartBNI.0026
Hart-046 HartXAI.0038/HartBNI.0026
Hart-038 HartXAI.0038/HartBNI.0026
Hart-021 HartXAI.0038/HartBNI.0026
Hart-013 HartXAI.0038/HartBNI.0026
Hart-015 HartXAI.0038/HartBNI.0026
Hart-028 HartXAI.0040/HartBNI.0026
Hart-035 HartXAI.0038/HartBNI.0026
Hart-016 HartXAI.0038/HartBNI.0026
Hart-017 HartXAI.0040/HartBNI.0026
Hart-036 HartXAI.0038/HartBNI.0026
Hart-014 HartXAI.0038/HartBNI.0026
Hart-037 HartXAI.0038/HartBNI.0026
Hart-004 HartXAI.0038/HartBNI.0026
Hart-024 HartXAI.0038/HartBNI.0026
Hart-033 HartXAI.0038/HartBNI.0026
Hart-020 HartXAI.0038/HartBNI.0026
Hart-027 HartXAI.0038/HartBNI.0026
Hart-019 HartXAI.0038/HartBNI.0026
Hart-012 HartXAI.0038/HartBNI.0026
Hart-011 HartXAI.0038/HartBNI.0026
Hart-009 HartXAI.0038/HartBNI.0026
Hart-005 HartXAI.0038/HartBNI.0026
Hart-006 HartXAI.0038/HartBNI.0026
0-6 hqSNVs
Hart-032 HartXAI.0043/
Hart-002 HartXAI.0039/HartBNI.0025
Hart-018 HartXAI.0039/HartBNI.0025
0 hqSNVs
Hart-045 HartXAI.0025/HartBNI.0010
Hart-044 HartXAI.0004/HartBNI.0010
Hart-042 HartXAI.0025/HartBNI.0010
Hart-022 HartXAI.0004/HartBNI.0028
Hart-041 HartXAI.0044/HartBNI.0032
Hart-023 HartXAI.0004/HartBNI.0010
Hart-029 HartXAI.0022/HartBNI.0031
Reference
Hart-030 HartXAI.0022/HartBNI.0031
Hart-040 HartXAI.0021/HartBNI.0007
Hart-025 HartXAI.0014/HartBNI.0030
Hart-026 HartXAI.0042/HartBNI.0016
Hart-031 HartXAI.0005/HartBNI.0016
Hart-043 HartXAI.0042/HartBNI.0016
0-1 hqSNVs
Hart-007 HartXAI.0041/HartBNI.0027
Hart-008 HartXAI.0041/HartBNI.0027
0 hqSNVs
193 hqSNVs

Figure 10: wgMLST-based tree of 45 *Salmonella* Hartford isolates from the time period surrounding the chia outbreak. Tree was built using character data on 5,039 loci. Each node is colour coded by PFGE-based clusters with non-cluster isolates appearing blue. WGS-based clusters are boxed with allele differences defining the cluster appearing beside the branch.



	PFGE-XbaI-pattern	PFGE-BlnI-pattern	SourceType
Hart-026	HartXAI.0042	HartBNI.0016	Human
Hart-031	HartXAI.0005	HartBNI.0016	Human
Hart-043	HartXAI.0042	HartBNI.0016	Human
Hart-029	HartXAI.0022	HartBNI.0031	Human
Hart-030	HartXAI.0022	HartBNI.0031	Human
Hart-040	HartXAI.0021	HartBNI.0007	Human
Hart-045	HartXAI.0025	HartBNI.0010	Human
Hart-044	HartXAI.0004	HartBNI.0010	Human
Hart-042	HartXAI.0025	HartBNI.0010	Human
Hart-022	HartXAI.0004	HartBNI.0028	Human
Hart-041	HartXAI.0044	HartBNI.0032	Human
Hart-023	HartXAI.0004	HartBNI.0010	Human
Hart-019	HartXAI.0038	HartBNI.0026	Human
Hart-020	HartXAI.0038	HartBNI.0026	Human
Hart-039	HartXAI.0038	HartBNI.0026	Human
Hart-046	HartXAI.0038	HartBNI.0026	Human
Hart-028	HartXAI.0040	HartBNI.0026	Human
Hart-035	HartXAI.0038	HartBNI.0026	Human
Hart-009	HartXAI.0038	HartBNI.0026	Human
Hart-027	HartXAI.0038	HartBNI.0026	Human
Hart-012	HartXAI.0038	HartBNI.0026	Human
Hart-005	HartXAI.0038	HartBNI.0026	Human
Hart-006	HartXAI.0038	HartBNI.0026	Human
Hart-038	HartXAI.0038	HartBNI.0026	Human
Hart-015	HartXAI.0038	HartBNI.0026	Human
Hart-033	HartXAI.0038	HartBNI.0026	Food
Hart-013	HartXAI.0038	HartBNI.0026	Human
Hart-011	HartXAI.0038	HartBNI.0026	Human
Hart-003	HartXAI.0038	HartBNI.0026	Human
Hart-017	HartXAI.0040	HartBNI.0026	Human
Hart-021	HartXAI.0038	HartBNI.0026	Human
Hart-010	HartXAI.0038	HartBNI.0026	Human
Hart-016	HartXAI.0038	HartBNI.0026	Human
Hart-036	HartXAI.0038	HartBNI.0026	Food
Hart-014	HartXAI.0038	HartBNI.0026	Human
Hart-037	HartXAI.0038	HartBNI.0026	Food
Hart-024	HartXAI.0038	HartBNI.0026	Food
Hart-001	HartXAI.0038	HartBNI.0026	Human
Hart-004	HartXAI.0038	HartBNI.0026	Human
Hart-002	HartXAI.0039	HartBNI.0025	Human
Hart-018	HartXAI.0039	HartBNI.0025	Human
Hart-032	HartXAI.0043		Human
Hart-007	HartXAI.0041	HartBNI.0027	Human
Hart-008	HartXAI.0041	HartBNI.0027	Human
Hart-025	HartXAI.0014	HartBNI.0030	Human

cases as it was identified after the outbreak investigation was closed on July 29th. No other PFGE-based clusters were identified from this serovar during this time period. The retrospective WGS analysis revealed three additional clusters from this serovar (Figures 9 and 10).

Table 7: Comparison of PFGE-based and WGS-based clusters of *Salmonella* Hartford isolates identified during the outbreak period.

Cluster Code	PFGE Analysis		WGS Analysis		
	Number of Isolates Ruled In	PFGE Patterns Included	Number of Isolates Ruled In	SNVPhyl Maximum Differences (hqSNVs)	wgMLST Maximum Differences (alleles)
PNC-Multi-1	26	HartXAI.0038/HartBNI.0026 HartXAI.0040/HartBNI.0026	27[1]	6	6

[1]Additional isolate was a PFGE match but excluded from outbreak as it was identified after the official outbreak investigation closed

3.2.2.2 Serovar Newport

A total of 161 of the 368 isolates sequenced were from *Salmonella* Newport. Phylogenetic trees on all *Salmonella* Newport isolates were produced. *Salmonella* Newport is considered to be a polyphyletic serovar with three well characterized global lineages in existence (67). These three lineages are separated by many thousands of hqSNVs or alleles from each other. Only *Salmonella* Newport II and III lineages had a significant amount of isolates circulating in Canada at this time, and just a single isolate from the *Salmonella* Newport I lineage was found in Canada during this time period. All 44 *Salmonella* Newport PNC-Multi-1 isolates, were from the *Salmonella* Newport III lineage.

A SNVPhyl tree was built for the comparison of all 120 *Salmonella* Newport III isolates using 6,730 hqSNVs across 92.4% of the reference genome, the fully closed *Salmonella* Newport isolate USDA-ARS-USMARC-1927 (downloaded December 2016) (Figure 11). Meanwhile, the wgMLST tree for these isolates was built using character data across 5,954 loci (Figure 12). The 44 *Salmonella* Newport PNC-Multi-1 isolates identified during the outbreak investigation were

split into two separate clusters by both WGS-based analysis methods. The two clusters were separated by almost 100 hqSNVs or alleles, and can be differentiated based on their size. The larger cluster contained 42 isolates and was separated by seven hqSNVs and 15 alleles (Table 8). This cluster included: all the human clinical cases initially identified; nineteen of the chia food isolates; and one additional isolate with a matching PFGE pattern that was not included in the official outbreak as it was found after the outbreak investigation was closed. The smaller cluster contains just three food isolates that had PFGE patterns with no human clinical matches. Isolates in this cluster have no hqSNV differences between them, but one allele difference via wgMLST analysis.

Four additional clusters had also been identified via PFGE analysis during this time period and WGS-based results agreed with PFGE-based analysis for two of these clusters, PNC-Newp-1[8] and PNC-Newp-2. Cluster PNC-Newp-3 consisted of three isolates; however, only two of the three isolates remained a cluster by WGS-based analysis, with the third isolate being excluded. Meanwhile cluster PNC-Newp-4 contained just two isolates; however, WGS-based analysis identified an additional two isolates with differing PFGE patterns that grouped with this cluster (Table 8). An additional eight clusters of two or more isolates were identified by WGS-based analysis from both analysis platforms, six of which were from the Newport III lineage (Figures 11 and 12).

[8] Isolates from this cluster are from the Newport II lineage and therefore not shown in Figures 13 and 14.

Figure 11: SNVPhyl tree of 120 *Salmonella* Newport III isolates from the time period surrounding the chia outbreak. Tree was built using 6,730 hqSNVs across 92.4% of the reference genome, the closed *Salmonella* Newport reference genome USDA-ARS-USMARC-1927. Each node is colour coded by PFGE-based clusters with non-cluster isolates appearing red. Node shape corresponds to the source of the isolate. WGS-based clusters are boxed with SNV differences defining the cluster appearing beside the cluster. Scale bar indicates the number of hqSNVs.

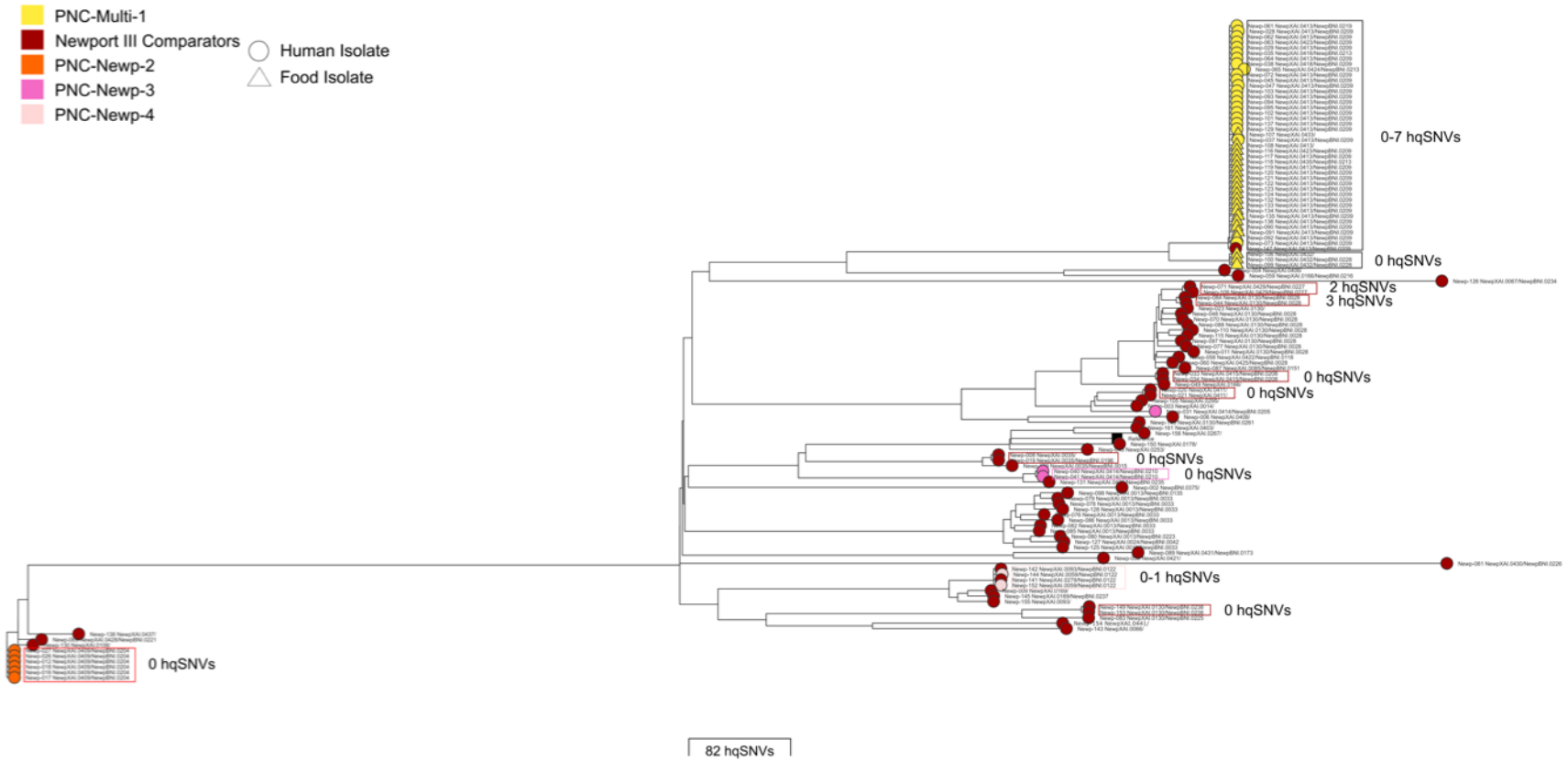

PNC-Multi-1
Newport III Comparators
PNC-Newp-2
PNC-Newp-3
PNC-Newp-4
Human Isolate
Food Isolate
0-7 hqSNVs
0 hqSNVs
2 hqSNVs
3 hqSNVs
0 hqSNVs
0 hqSNVs
0 hqSNVs
0 hqSNVs
0-1 hqSNVs
0 hqSNVs
0 hqSNVs
82 hqSNVs

Figure 12: wgMLST-based tree of 120 *Salmonella* Newport III isolates from the time period surrounding the chia outbreak. Tree was built using character data on 5,954 loci. Each node is colour coded by PFGE-based clusters with non-cluster isolates appearing red. WGS-based clusters are boxed with allele differences defining the cluster appearing beside the branch.

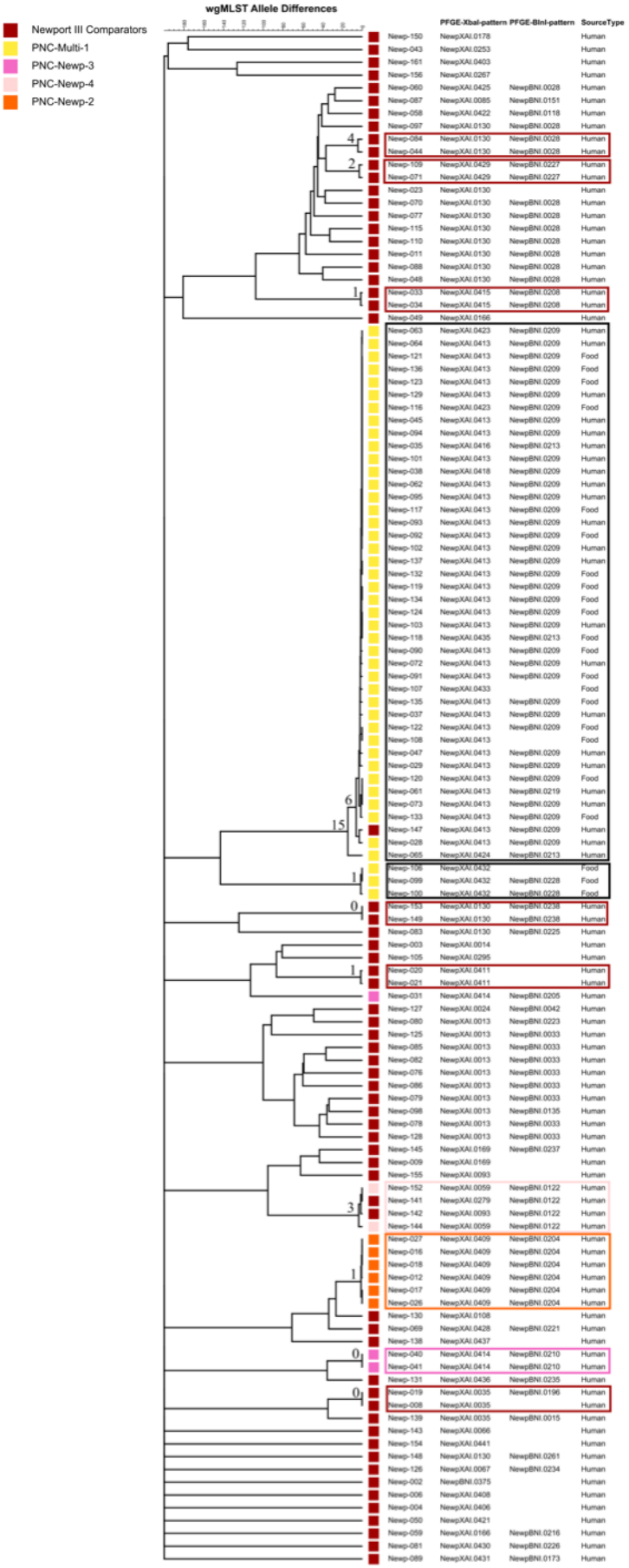

wgMLST Allele Differences
Newport III Comparators
PNC-Multi-1
PNC-Newp-3
PNC-Newp-4
PNC-Newp-2
PFGE-XbaI-pattern
PFGE-BlnI-pattern
SourceType

Table 8: Comparison of PFGE-based and WGS-based clusters of *Salmonella* Newport isolates identified during the outbreak period.

Cluster Code	PFGE Analysis		WGS Analysis		
	Number of Isolates Ruled In	PFGE Patterns Included	Number of Isolates Ruled In	SNVPhyl Maximum Differences (hqSNVs)	wgMLST Maximum Differences (alleles)
PNC-Multi-1	44	NewpXAI.0413 NewpXAI.0416 NewpXAI.0418 NewpXAI.0423 NewpXAI.0424	42[2]	7	15
		NewpXAI.0432[1] NewpXAI.0433[1] NewpXAI.0435[1]	3[1]	0	1
PNC-Newp-1	6	NewpXAI.0015	6	2	3
PNC-Newp-2	6	NewpXAI.0409/NewpBNI.0204	6	0	0
PNC-Newp-3	3	NewpXAI.0414/NewpBNI.0210 NewpXAI.0414/NewpBNI.0205	2[3]	0	0
PNC-Newp-4	2	NewpXAI.0059/NewpBNI.0122	4[4]	1	3

[1]No clinical matches found.
[2]Additional isolate was a PFGE match but excluded from outbreak as it was identified after the official outbreak investigation closed.
[3]Isolate ruled out by WGS analysis had a unique, but highly similar, *Bln*1 pattern.
[4]Additional isolates ruled in by WGS analysis had unique *Xba*1 patterns, but the same *Bln*1 pattern as PFGE cluster.

3.2.2.3 Serovar Oranienburg

A total of 66 of the 368 isolates sequenced were from serovar Oranienburg. Phylogenetic trees were produced using both SNVPhyl and wgMLST, and it was noted that *Salmonella* Oranienburg was a highly diverse and polyphyletic serovar with multiple lineages circulating in Canada that were separated by many thousands of SNVs and alleles. However, the majority (n=48) of isolates were from a single lineage. A SNVPhyl tree was built on this group of isolates and was constructed using 2,196 hqSNVs across 93.1% of the reference genome, the SPAdes assembled isolate Orani-024 (Figure 13). A dendrogram was also produced from the wgMLST

data for this group of isolates and was built using data on 5,100 loci (Figure 14).

Included in the main lineage of *Salmonella* Oranienburg was five of the seven *Salmonella* Oranienburg isolates from the PNC-Multi-1 outbreak investigation. Only two of these five isolates, both human clinical cases, formed a cluster by WGS-based analysis that was separated by ten hqSNVs or alleles (Table 9). These two clinical cases had PFGE patterns that were included in the outbreak investigation due to a match to a food isolate collected in the US, which we were unable to acquire in this investigation. The other three *Salmonella* Oranienburg isolates from PNC-Multi-1 that were within this main clade were very diverse (separated by >140 hqSNVs or >125 alleles), and are therefore unrelated by WGS. (Figures 13 and 14) The two other *Salmonella* Oranienburg isolates implicated in PNC-Multi-1 were two food sample isolates with PFGE patterns that did not produce any clinical cases. These isolates were indistinguishable by WGS analysis (Table 9) and clustered well away (>14,000 hqSNVs and >3,200 alleles) from any comparator isolates.

Table 9: Comparison of PFGE-based and WGS-based clusters of *Salmonella* Oranienburg isolates identified during the outbreak period.

Cluster Code	PFGE Analysis		WGS Analysis		
	Number of Isolates Ruled In	PFGE Patterns Included	Number of Isolates Ruled In	SNVPhyl Maximum Differences (hqSNVs)	wgMLST Maximum Differences (alleles)
PNC-Multi-1	7	OraniXAI.0005/OraniBNI.0007 OraniXAI.0005/OraniBNI.0073 OraniXAI.0005/OraniBNI.0075	2	10	10
		OraniXAI.0006/OraniBNI.0003 OraniXAI.0144/OraniBNI.0083[1]	2[1]	0	0
PNC-Orani-1	3	OraniXAI.0005/OraniBNI.0007	0	0	0

[1]No clinical matches found.

One additional PFGE-based cluster of three *Salmonella* Oranienburg isolates was identified during this time, denoted PNC-Orani-1. All PNC-Orani-1 isolates were found in the

Figure 13: SNVPhyl tree of 48 *Salmonella* Oranienburg isolates from the main Oranienburg lineage circulating in Canada during the time period surrounding the chia outbreak. Tree was built using 2,196 hqSNVs across 93.1% of the reference genome, the SPAdes assembled genome Orani-024. Each node is colour coded by PFGE-based cluster with non-cluster isolates appearing purple. Node shape corresponds to the source of the isolate. WGS-based clusters are boxed with SNV differences defining the cluster beside the cluster. Scale bar indicates the number of hqSNVs.

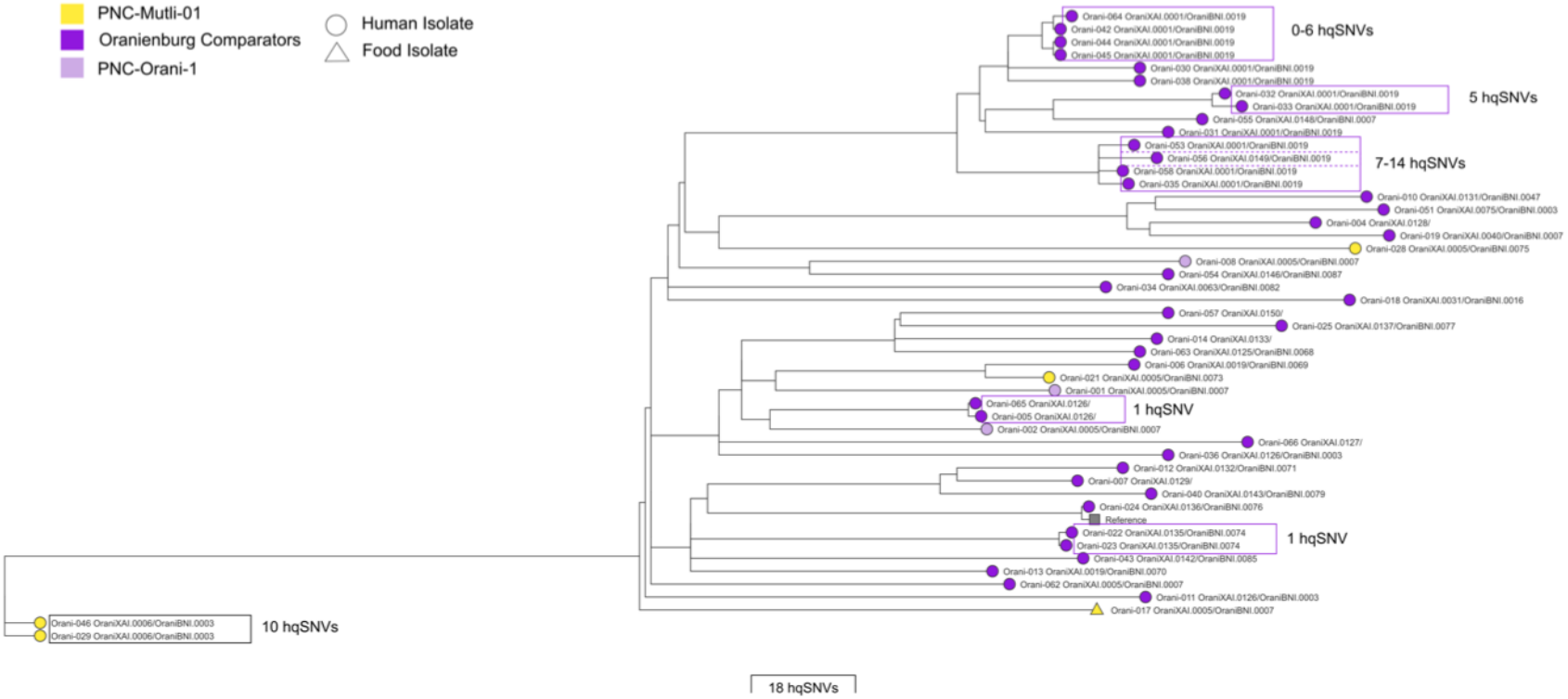
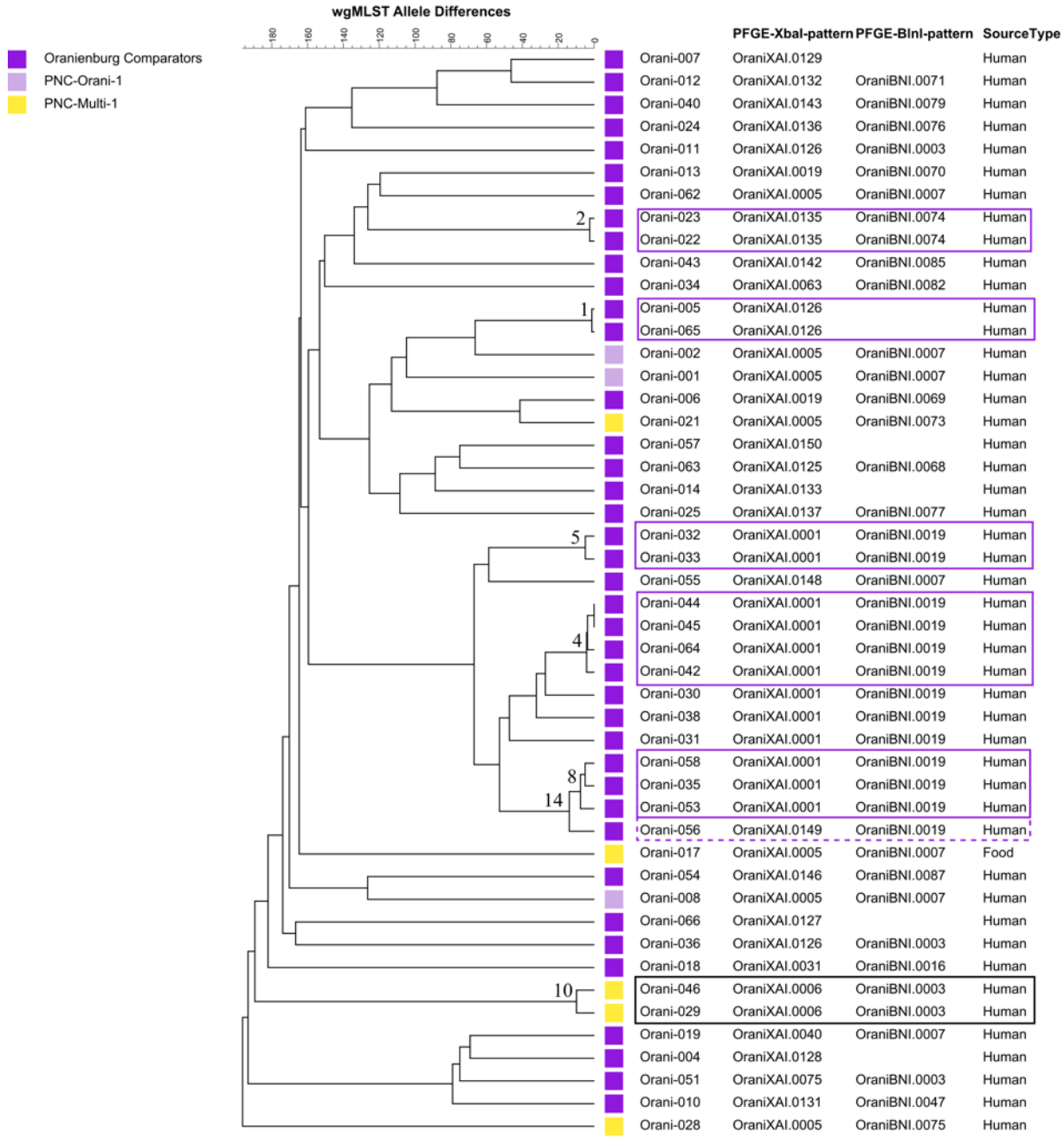

PNC-Mutli-01
Oranienburg Comparators
PNC-Orani-1
Human Isolate
Food Isolate
Orani-064 OraniXAI.0001/OraniBNI.0019
Orani-042 OraniXAI.0001/OraniBNI.0019
Orani-044 OraniXAI.0001/OraniBNI.0019
Orani-045 OraniXAI.0001/OraniBNI.0019
0-6 hqSNVs
Orani-030 OraniXAI.0001/OraniBNI.0019
Orani-038 OraniXAI.0001/OraniBNI.0019
Orani-032 OraniXAI.0001/OraniBNI.0019
Orani-033 OraniXAI.0001/OraniBNI.0019
5 hqSNVs
Orani-055 OraniXAI.0148/OraniBNI.0007
Orani-031 OraniXAI.0001/OraniBNI.0019
Orani-053 OraniXAI.0001/OraniBNI.0019
Orani-056 OraniXAI.0149/OraniBNI.0019
Orani-058 OraniXAI.0001/OraniBNI.0019
Orani-035 OraniXAI.0001/OraniBNI.0019
7-14 hqSNVs
Orani-010 OraniXAI.0131/OraniBNI.0047
Orani-051 OraniXAI.0075/OraniBNI.0003
Orani-004 OraniXAI.0128/
Orani-019 OraniXAI.0040/OraniBNI.0007
Orani-028 OraniXAI.0005/OraniBNI.0075
Orani-008 OraniXAI.0005/OraniBNI.0007
Orani-054 OraniXAI.0146/OraniBNI.0087
Orani-034 OraniXAI.0063/OraniBNI.0082
Orani-018 OraniXAI.0031/OraniBNI.0016
Orani-057 OraniXAI.0150/
Orani-025 OraniXAI.0137/OraniBNI.0077
Orani-014 OraniXAI.0133/
Orani-063 OraniXAI.0125/OraniBNI.0068
Orani-006 OraniXAI.0019/OraniBNI.0069
Orani-021 OraniXAI.0005/OraniBNI.0073
Orani-001 OraniXAI.0005/OraniBNI.0007
Orani-065 OraniXAI.0126/
Orani-005 OraniXAI.0126/
1 hqSNV
Orani-002 OraniXAI.0005/OraniBNI.0007
Orani-066 OraniXAI.0127/
Orani-036 OraniXAI.0126/OraniBNI.0003
Orani-012 OraniXAI.0132/OraniBNI.0071
Orani-007 OraniXAI.0129/
Orani-040 OraniXAI.0143/OraniBNI.0079
Orani-024 OraniXAI.0136/OraniBNI.0076
Reference
Orani-022 OraniXAI.0135/OraniBNI.0074
Orani-023 OraniXAI.0135/OraniBNI.0074
1 hqSNV
Orani-043 OraniXAI.0142/OraniBNI.0085
Orani-013 OraniXAI.0019/OraniBNI.0070
Orani-062 OraniXAI.0005/OraniBNI.0007
Orani-011 OraniXAI.0126/OraniBNI.0003
Orani-017 OraniXAI.0005/OraniBNI.0007
Orani-046 OraniXAI.0006/OraniBNI.0003
Orani-029 OraniXAI.0006/OraniBNI.0003
10 hqSNVs
18 hqSNVs

Figure 14: wgMLST-based tree of 48 *Salmonella* Oranienburg isolates from the main Oranienburg lineage circulating in Canada during the time period surrounding the chia outbreak. Tree was built using character data on 5,100 loci. Each node is colour coded by PFGE-based clusters with non-cluster isolates appearing purple. WGS identified clusters are boxed with allele differences defining the cluster appearing beside the branch.



	PFGE-XbaI-pattern	PFGE-BlnI-pattern	SourceType
Orani-007	OraniXAI.0129		Human
Orani-012	OraniXAI.0132	OraniBNI.0071	Human
Orani-040	OraniXAI.0143	OraniBNI.0079	Human
Orani-024	OraniXAI.0136	OraniBNI.0076	Human
Orani-011	OraniXAI.0126	OraniBNI.0003	Human
Orani-013	OraniXAI.0019	OraniBNI.0070	Human
Orani-062	OraniXAI.0005	OraniBNI.0007	Human
Orani-023	OraniXAI.0135	OraniBNI.0074	Human
Orani-022	OraniXAI.0135	OraniBNI.0074	Human
Orani-043	OraniXAI.0142	OraniBNI.0085	Human
Orani-034	OraniXAI.0063	OraniBNI.0082	Human
Orani-005	OraniXAI.0126		Human
Orani-065	OraniXAI.0126		Human
Orani-002	OraniXAI.0005	OraniBNI.0007	Human
Orani-001	OraniXAI.0005	OraniBNI.0007	Human
Orani-006	OraniXAI.0019	OraniBNI.0069	Human
Orani-021	OraniXAI.0005	OraniBNI.0073	Human
Orani-057	OraniXAI.0150		Human
Orani-063	OraniXAI.0125	OraniBNI.0068	Human
Orani-014	OraniXAI.0133		Human
Orani-025	OraniXAI.0137	OraniBNI.0077	Human
Orani-032	OraniXAI.0001	OraniBNI.0019	Human
Orani-033	OraniXAI.0001	OraniBNI.0019	Human
Orani-055	OraniXAI.0148	OraniBNI.0007	Human
Orani-044	OraniXAI.0001	OraniBNI.0019	Human
Orani-045	OraniXAI.0001	OraniBNI.0019	Human
Orani-064	OraniXAI.0001	OraniBNI.0019	Human
Orani-042	OraniXAI.0001	OraniBNI.0019	Human
Orani-030	OraniXAI.0001	OraniBNI.0019	Human
Orani-038	OraniXAI.0001	OraniBNI.0019	Human
Orani-031	OraniXAI.0001	OraniBNI.0019	Human
Orani-058	OraniXAI.0001	OraniBNI.0019	Human
Orani-035	OraniXAI.0001	OraniBNI.0019	Human
Orani-053	OraniXAI.0001	OraniBNI.0019	Human
Orani-056	OraniXAI.0149	OraniBNI.0019	Human
Orani-017	OraniXAI.0005	OraniBNI.0007	Food
Orani-054	OraniXAI.0146	OraniBNI.0087	Human
Orani-008	OraniXAI.0005	OraniBNI.0007	Human
Orani-066	OraniXAI.0127		Human
Orani-036	OraniXAI.0126	OraniBNI.0003	Human
Orani-018	OraniXAI.0031	OraniBNI.0016	Human
Orani-046	OraniXAI.0006	OraniBNI.0003	Human
Orani-029	OraniXAI.0006	OraniBNI.0003	Human
Orani-019	OraniXAI.0040	OraniBNI.0007	Human
Orani-004	OraniXAI.0128		Human
Orani-051	OraniXAI.0075	OraniBNI.0003	Human
Orani-010	OraniXAI.0131	OraniBNI.0047	Human
Orani-028	OraniXAI.0005	OraniBNI.0075	Human

main lineage (Figures 15 and 16), and were highly distinct from each other by both WGS-based analyses methods and would therefore not be considered a cluster by WGS-based analysis (Table 9). WGS-based analysis of all lineages of *Salmonella* Oranienburg revealed an additional six clusters of two or more isolates separated by less than 10 SNVs or alleles. Of note five of these WGS-based clusters were within the main lineage of *Salmonella* Oranienburg isolates circulating in Canada (Figures 15 and 16).

3.2.2.4 Serovar Saintpaul

A total of 94 of the 368 isolates sequenced were from serovar Saintpaul. Phylogenetic trees were produced using both SNVPhyl and wgMLST from these isolates, and it was noted that this serovar was also polyphyletic serovar with multiple lineages circulating in Canada. A single dominant lineage of *Salmonella* Saintpaul isolates was noted to be circulating in Canada that consisted of 63 isolates, including all 26 isolates identified as part of the PNC-Multi-1 outbreak investigation. A SNVPhyl tree was built on this main lineage of *Salmonella* Saintpaul isolates using 1,279 hqSNVs across 93.25% of the reference genome, the SPAdes assembled genome Saint-012 (Figure 15). A wgMLST dendrogram was also built for this group of isolates for comparison and was produced using data from 4,766 loci (Figure 16).

The 26 PNC-Multi-1 isolates from serovar Saintpaul, were split into two clusters by both WGS-based analysis methods. These clusters were separated by 75 hqSNVs or 70 alleles. The larger of the two clusters consists of 22 of the 26 isolates identified in the initial outbreak investigation plus an additional four isolates initially excluded from the outbreak. These additional isolates included two isolates that had matching PFGE patterns to the outbreak but were identified after the official outbreak investigation was closed, and therefore were excluded on that reason. The other two isolates were excluded from the outbreak as they had different

PFGE patterns from the outbreak patterns; however, these isolates would have been included in the outbreak using a WGS-based analysis. The smaller cluster of chia outbreak isolates consisted of four food isolates that had a unique PFGE pattern that was not in the official case definition, as there were no clinical case matches. However, one PFGE matching clinical case was found after the official outbreak investigation was closed, and this isolate also clustered with these four isolates by WGS-based analysis (Table 10).

Table 10: Comparison of PFGE-based and WGS-based clusters of *Salmonella* Saintpaul isolates identified during the outbreak period.

Cluster Code	PFGE Analysis		WGS Analysis		
	Number of Isolates Ruled In	PFGE Patterns Included	Number of Isolates Ruled In	SNVPhyl Maximum Differences (hqSNVs)	wgMLST Maximum Differences (alleles)
PNC-Multi-1	25	SainXAI.0005/SainBNI.00010 SainXAI.0005/SainBNI.00073 SainXAI.0170/SainBNI.0081[1]	25[2,3]	12	2
			5[2]	0	0
PNC-Sain-1	14	SainXAI.0005 SainXAI.0005/SainBNI.0005 SainXAI.0159/SainBNI.0005	4[3,4]	1	2
			2	1	1
PNC-Sain-2	4	SainXAI.0046/SainBNI.0088 SainXAI.0046	4	1	1

[1]No Clinical Matches.
[2]Additional isolates identified were initially excluded from cluster due to identification past the official date the investigation closed.
[3]Additional isolate identified were initially excluded from cluster due to unique PFGE patterns.
[4]Additional isolate identified was initially excluded from cluster due to temporal distance.

Two other PFGE-based clusters, identified by PulseNet Canada, were from serovar Saintpaul and were identified during this time period. One of these clusters, PNC-Sain-1, was found within the main lineage of *Salmonella* Saintpaul and consisted of 14 isolates; however, both WGS-based analysis methods fully resolved this cluster. Four isolates from PNC-Sain-1 did go on to form two separate WGS-based clusters. The first of which consisted of just two isolates separated by 2 hqSNVs and 1 allele. The second cluster contained the other two PNC-Sain-1

Figure 15: SNVPhyl tree of 63 *Salmonella* Saintpaul isolates from the main Saintpaul lineage circulating in Canada during the time period surround the chia outbreak. Tree was built using 1,279 hqSNVs across 93.25% of the reference genome, the SPAdes assembled genome Saint-012. Each node is colour coded PFGE-based cluster with non-cluster isolates appearing green. Node shape corresponds to the source of the isolate. WGS-based clusters are boxed with SNV differences defining the cluster beside the cluster. Scale bar indicates the number of hqSNVs.

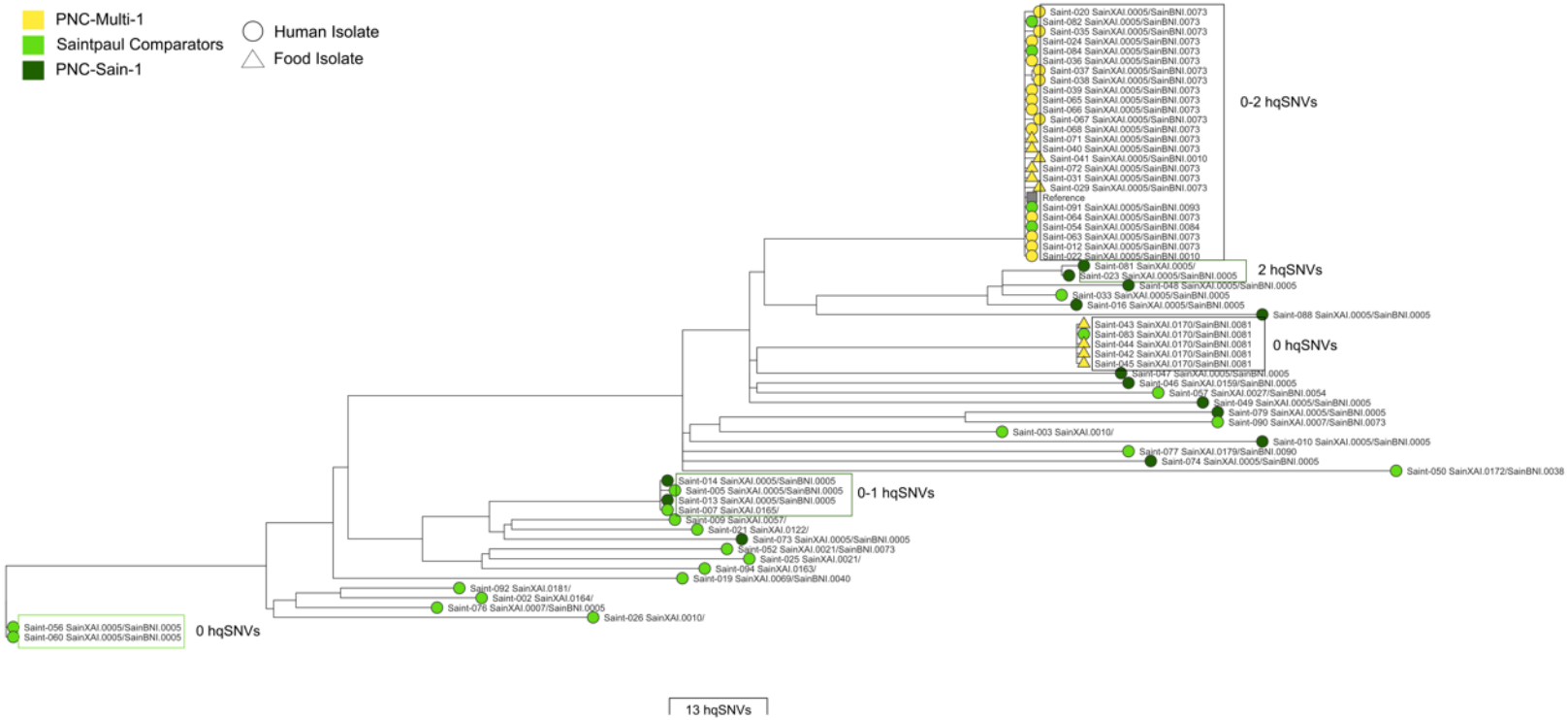
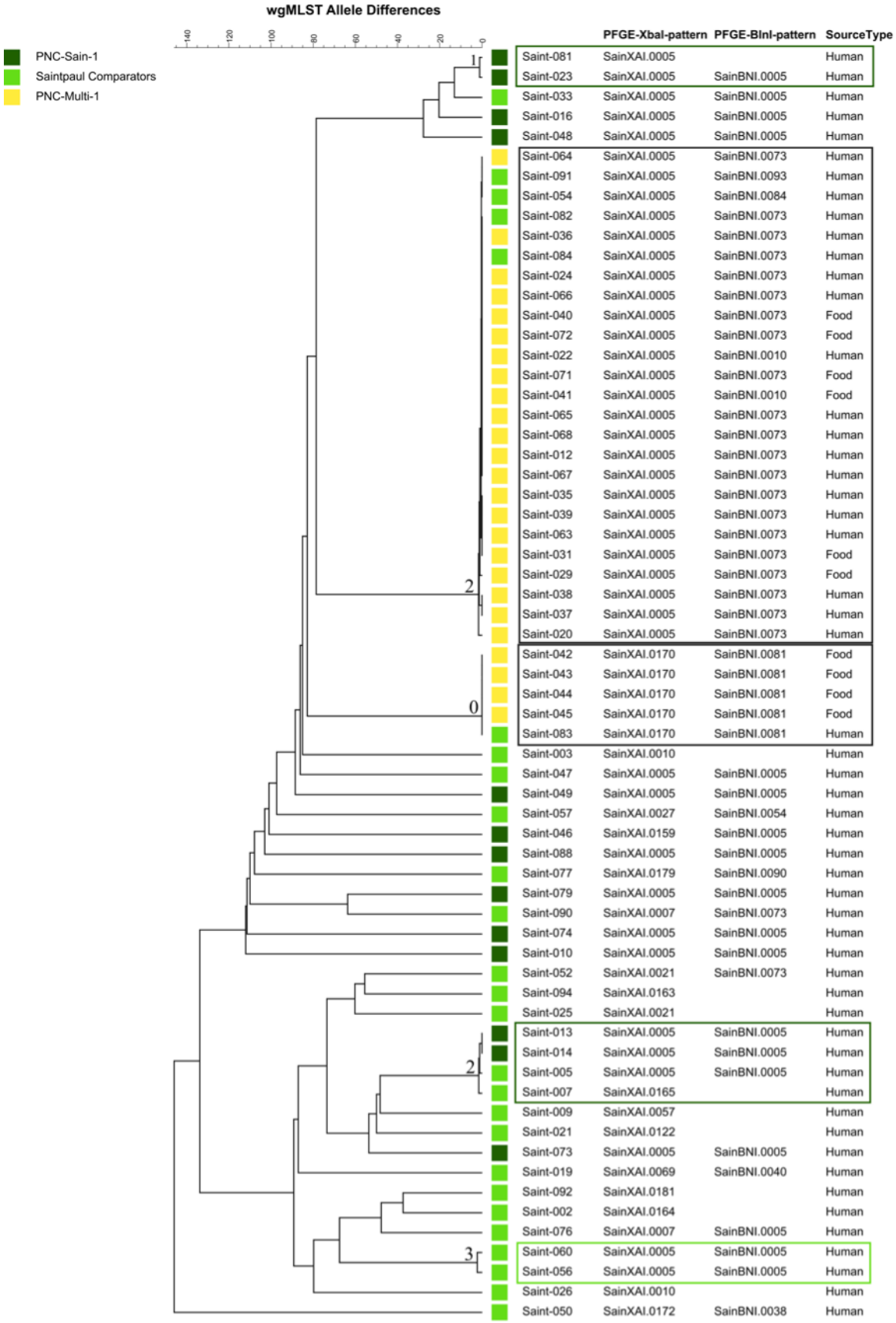

PNC-Multi-1
Saintpaul Comparators
PNC-Sain-1
Human Isolate
Food Isolate
Saint-020 SainXAI.0005/SainBNI.0073
Saint-082 SainXAI.0005/SainBNI.0073
Saint-035 SainXAI.0005/SainBNI.0073
Saint-024 SainXAI.0005/SainBNI.0073
Saint-084 SainXAI.0005/SainBNI.0073
Saint-036 SainXAI.0005/SainBNI.0073
Saint-037 SainXAI.0005/SainBNI.0073
Saint-038 SainXAI.0005/SainBNI.0073
Saint-039 SainXAI.0005/SainBNI.0073
Saint-065 SainXAI.0005/SainBNI.0073
Saint-066 SainXAI.0005/SainBNI.0073
Saint-067 SainXAI.0005/SainBNI.0073
Saint-068 SainXAI.0005/SainBNI.0073
Saint-071 SainXAI.0005/SainBNI.0073
Saint-040 SainXAI.0005/SainBNI.0073
Saint-041 SainXAI.0005/SainBNI.0010
Saint-072 SainXAI.0005/SainBNI.0073
Saint-031 SainXAI.0005/SainBNI.0073
Saint-029 SainXAI.0005/SainBNI.0073
Reference
Saint-091 SainXAI.0005/SainBNI.0093
Saint-064 SainXAI.0005/SainBNI.0073
Saint-054 SainXAI.0005/SainBNI.0084
Saint-063 SainXAI.0005/SainBNI.0073
Saint-012 SainXAI.0005/SainBNI.0073
Saint-022 SainXAI.0005/SainBNI.0010
0-2 hqSNVs
Saint-081 SainXAI.0005/
Saint-023 SainXAI.0005/SainBNI.0005
Saint-048 SainXAI.0005/SainBNI.0005
Saint-033 SainXAI.0005/SainBNI.0005
Saint-016 SainXAI.0005/SainBNI.0005
2 hqSNVs
Saint-088 SainXAI.0005/SainBNI.0005
Saint-043 SainXAI.0170/SainBNI.0081
Saint-083 SainXAI.0170/SainBNI.0081
Saint-044 SainXAI.0170/SainBNI.0081
Saint-042 SainXAI.0170/SainBNI.0081
Saint-045 SainXAI.0170/SainBNI.0081
0 hqSNVs
Saint-047 SainXAI.0005/SainBNI.0005
Saint-046 SainXAI.0159/SainBNI.0005
Saint-057 SainXAI.0027/SainBNI.0054
Saint-049 SainXAI.0005/SainBNI.0005
Saint-079 SainXAI.0005/SainBNI.0005
Saint-090 SainXAI.0007/SainBNI.0073
Saint-003 SainXAI.0010/
Saint-010 SainXAI.0005/SainBNI.0005
Saint-077 SainXAI.0179/SainBNI.0090
Saint-074 SainXAI.0005/SainBNI.0005
Saint-050 SainXAI.0172/SainBNI.0038
Saint-014 SainXAI.0005/SainBNI.0005
Saint-005 SainXAI.0005/SainBNI.0005
Saint-013 SainXAI.0005/SainBNI.0005
Saint-007 SainXAI.0165/
Saint-009 SainXAI.0057/
Saint-021 SainXAI.0122/
Saint-073 SainXAI.0005/SainBNI.0005
Saint-052 SainXAI.0021/SainBNI.0073
Saint-025 SainXAI.0021/
Saint-094 SainXAI.0163/
Saint-019 SainXAI.0069/SainBNI.0040
0-1 hqSNVs
Saint-092 SainXAI.0181/
Saint-002 SainXAI.0164/
Saint-076 SainXAI.0007/SainBNI.0005
Saint-026 SainXAI.0010/
Saint-056 SainXAI.0005/SainBNI.0005
Saint-060 SainXAI.0005/SainBNI.0005
0 hqSNVs
13 hqSNVs

Figure 16: wgMLST-based tree of 63 *Salmonella* Saintpaul isolates from the main Saintpaul lineage circulating in Canada during the time period surrounding the chia outbreak. Tree was built using character data on 5,100 loci. Each node is colour coded by PFGE-based cluster with non-cluster isolates appearing green. WGS-based clusters are boxed with allele differences defining the cluster appearing beside the branch.



	PFGE-XbaI-pattern	PFGE-BlnI-pattern	SourceType
Saint-081	SainXAI.0005		Human
Saint-023	SainXAI.0005	SainBNI.0005	Human
Saint-033	SainXAI.0005	SainBNI.0005	Human
Saint-016	SainXAI.0005	SainBNI.0005	Human
Saint-048	SainXAI.0005	SainBNI.0005	Human
Saint-064	SainXAI.0005	SainBNI.0073	Human
Saint-091	SainXAI.0005	SainBNI.0093	Human
Saint-054	SainXAI.0005	SainBNI.0084	Human
Saint-082	SainXAI.0005	SainBNI.0073	Human
Saint-036	SainXAI.0005	SainBNI.0073	Human
Saint-084	SainXAI.0005	SainBNI.0073	Human
Saint-024	SainXAI.0005	SainBNI.0073	Human
Saint-066	SainXAI.0005	SainBNI.0073	Human
Saint-040	SainXAI.0005	SainBNI.0073	Food
Saint-072	SainXAI.0005	SainBNI.0073	Food
Saint-022	SainXAI.0005	SainBNI.0010	Human
Saint-071	SainXAI.0005	SainBNI.0073	Food
Saint-041	SainXAI.0005	SainBNI.0010	Food
Saint-065	SainXAI.0005	SainBNI.0073	Human
Saint-068	SainXAI.0005	SainBNI.0073	Human
Saint-012	SainXAI.0005	SainBNI.0073	Human
Saint-067	SainXAI.0005	SainBNI.0073	Human
Saint-035	SainXAI.0005	SainBNI.0073	Human
Saint-039	SainXAI.0005	SainBNI.0073	Human
Saint-063	SainXAI.0005	SainBNI.0073	Human
Saint-031	SainXAI.0005	SainBNI.0073	Food
Saint-029	SainXAI.0005	SainBNI.0073	Food
Saint-038	SainXAI.0005	SainBNI.0073	Human
Saint-037	SainXAI.0005	SainBNI.0073	Human
Saint-020	SainXAI.0005	SainBNI.0073	Human
Saint-042	SainXAI.0170	SainBNI.0081	Food
Saint-043	SainXAI.0170	SainBNI.0081	Food
Saint-044	SainXAI.0170	SainBNI.0081	Food
Saint-045	SainXAI.0170	SainBNI.0081	Food
Saint-083	SainXAI.0170	SainBNI.0081	Human
Saint-003	SainXAI.0010		Human
Saint-047	SainXAI.0005	SainBNI.0005	Human
Saint-049	SainXAI.0005	SainBNI.0005	Human
Saint-057	SainXAI.0027	SainBNI.0054	Human
Saint-046	SainXAI.0159	SainBNI.0005	Human
Saint-088	SainXAI.0005	SainBNI.0005	Human
Saint-077	SainXAI.0179	SainBNI.0090	Human
Saint-079	SainXAI.0005	SainBNI.0005	Human
Saint-090	SainXAI.0007	SainBNI.0073	Human
Saint-074	SainXAI.0005	SainBNI.0005	Human
Saint-010	SainXAI.0005	SainBNI.0005	Human
Saint-052	SainXAI.0021	SainBNI.0073	Human
Saint-094	SainXAI.0163		Human
Saint-025	SainXAI.0021		Human
Saint-013	SainXAI.0005	SainBNI.0005	Human
Saint-014	SainXAI.0005	SainBNI.0005	Human
Saint-005	SainXAI.0005	SainBNI.0005	Human
Saint-007	SainXAI.0165		Human
Saint-009	SainXAI.0057		Human
Saint-021	SainXAI.0122		Human
Saint-073	SainXAI.0005	SainBNI.0005	Human
Saint-019	SainXAI.0069	SainBNI.0040	Human
Saint-092	SainXAI.0181		Human
Saint-002	SainXAI.0164		Human
Saint-076	SainXAI.0007	SainBNI.0005	Human
Saint-060	SainXAI.0005	SainBNI.0005	Human
Saint-056	SainXAI.0005	SainBNI.0005	Human
Saint-026	SainXAI.0010		Human
Saint-050	SainXAI.0172	SainBNI.0038	Human

isolates, plus an additional two *Salmonella* Saintpaul isolates. The other PFGE-based cluster, PNC-Sain-2, was located in a separate lineage and both WGS-based analysis methods confirmed this cluster (Table 10). Lastly, both WGS-based methods identified two additional clusters, one of which resided within the main lineage of *Salmonella* Saintpaul (Figures 15 and 16).

Chapter 4: Discussion

Increasingly, laboratories performing foodborne disease surveillance are looking towards WGS to provide a high-resolution solution to their surveillance mandates (69). This is because WGS has shown the ability to provide improved resolution for surveillance activities at reduced costs when compared to traditional subtyping methodologies (85). For example, the application of WGS for nationwide *Listeria monocytogenes* outbreak detection and investigation in the United States has not only increased the detection of disease clusters but has also more frequently solved outbreaks when combined with the available epidemiological data in comparison to previous years that relied on the traditional molecular subtyping methods. These results were so transformative that WGS for routine surveillance is being implemented for other enteric pathogens in the US, including *Salmonella* (84). Apart from its application in outbreak detection and investigation, WGS has the potential to provide a cost-effective way to answer several additional questions about an isolate under study. This not only includes the potential of replacing other subtyping techniques, such as serotyping via *in silico* tests (41, 42, 77), but could also be used in other contexts such as providing information for risk assessments (69) and on antibiotic resistances (69, 108). However, before WGS can be fully utilized, the full validation of this technology and a complete understanding of its impacts on existing surveillance systems is required.

4.1 WGS-Based Prediction of *Salmonella* Serovars

4.1.1 Initial SISTR Validation Uncovered Four Areas of Improvement

The initial validation of serovar prediction using SISTR resulted in a direct match to the phenotypic serovar for 69.7% of the 492 isolates tested. The results from this initial validation were further explored and used to develop a series of recommendations to refine and improve upon SISTR's ability to provide an *in silico* serovar prediction. The recommendations that were

developed could be broken down into four main categories: improvements for O-antigen prediction; recognition of incompatible results; expansion and curation of the cgMLST database; and curation of the flagellar antigen databases.

4.1.1.1 Improving O-Antigen Prediction with Changes to Code and Library Preparation

In our initial validation, over 13% of the isolates tested returned without an O-antigen prediction, and presented a major issue in SISTR, as no serovar prediction was made for these isolates. At the time of the initial validation, the detection of an O-antigen was considered a key quality metric. Therefore, when no O-antigen was detected SISTR did not provide a full serovar prediction, even if there was sufficient evidence from other sources of information to make a serovar call (Personal Communication, SISTR Development Team, 2015). Two separate issues were uncovered in the exploration into a lack of O-antigen prediction, and the developmental and procedural fixes for both ultimately improved O-antigen predictions in SISTR.

The first issue uncovered was seen in 20% of the isolates that returned without an O-antigen prediction. All isolates in this group were from a single somatic serogroup, specifically the C2-C3 group, and the gene sequences that SISTR uses to make the C2-C3 call were present within the genomic assemblies. This indicated a technical issue was preventing the call, and the SISTR developers uncovered a misplaced period within the prediction algorithm for this serogroup, leading to the non-call for these isolates. The second issue uncovered was responsible for the remaining 80% of isolates that lacked an O-antigen prediction. For these isolates, the *wzx* and *wzy* genes that SISTR uses to make an O-antigen prediction were missing from not only the assemblies, but also the raw read data. The missing *rfb* operon sequencing data was ultimately attributed to a size selection step in the preparation of genomic libraries, which was filtering out short fragments.

The preparation of genomic libraries is an important step in WGS protocols, and any biases introduced in this part of the method will go on to influence and affect all downstream results. It has been well established that the mechanical shearing of genomic DNA is the best method for fragmenting the genomic material used in sequencing (109-111). While mechanical shearing through sonication-based methods provides an unbiased fragmentation of DNA, it does display some drawbacks, as the equipment required for this process is expensive and these processes also increase the length of the library preparation steps. On the other hand, enzymatic fragmentation utilizes unique transposases to fragment DNA at specific sequence motifs and this not only reduces the need for expensive infrastructure but also combines multiple steps into a single reaction, shortening the length of the protocol (109). Previous reports have shown that the various enzymatic fragmentation kits on the market, including the Nextera XT kit used in this study, show some bias in the fragmentation of the DNA, leading to areas of low coverage (109-111) and more variable median insert sizes (109, 111). Biases introduced in this step would further be exacerbated following any size selection of genomic libraries, which was done in this study to optimize sequencing (112).

The size selected data showed a lower average base coverage, a greater number of non-covered bases, and a lower percentage of bases under 30X coverage over the entire *rfb* region (Table 2), in comparison to the non-size selected data. While the data generated from size selected libraries is optimized to ensure near universal maximum read length and sufficient distance between read pairs (112), it also lead to the loss of data over regions of increased fragmentation, and in the case of this study left out key pieces of information for the *in silico* serotyping of *Salmonella* isolates. It is unlikely this same loss of data would occur with mechanically fragmented genomic DNA that underwent a size selection step, due to its unbiased

approach to fragmentation.

At this time, the use of mechanical fragmentation techniques by Canada's national diagnostic and reference laboratories with surveillance functions is not feasible. The acquisition of equipment for the mechanical fragmentation of genomic DNA is unlikely due to the high infrastructural investments costs and the faster preparation time of enzymatic kits continue to make them the preferred methodology going forward (Personal Communication, Celine Nadon, 2016). However, the potential for alternative fragmentation kits that display reduced biases could be explored in the future. Recent data has suggested that the KAPA HyperPlus kit shows less bias in its fragmentation in comparison to the Nextera XT kit, is more efficient in adaptor ligation, and still maintains the preparation speed that makes enzymatic kits desirable (111). An investigation into the KAPA HyperPlus kit represents a future possible direction; however, in the meantime, the suspension of size selection on Nextera XT libraries of *Salmonella* is recommended to ensure *in silico* serotyping data is complete.

All data generated in the *in silico* serotyping validation study was produced from size selected libraries. To improve serovar predictions in these cases, and any future cases where an O-antigen is not found, it was proposed that the correlation between an H-antigen prediction and a cgMLST serovar prediction was sufficient to provide a serovar call. Algorithmic logic was therefore added into the SISTR program that produced a long list of potential serovars based solely on flagellar antigens detected when no O-antigen was found, and a serovar prediction was then made from this list using cgMLST results.

4.1.1.2 Importance of Recognizing Genotypic Matches

The initial validation of SISTR uncovered a set of isolates that had incompatible phenotypic and *in silico* serotyping designations. These incompatibilities specifically resulted

from the carriage of unexpressed antigenic determinants and were reported in 6.5% of the isolates tested. The use of a WGS-based analysis platform for serotyping represents a paradigm shift in the understanding of serotypes. In these tests, the genomic carriage of antigenic markers defines the serovar, not the phenotypic expression of those markers. On one level this new paradigm presents the opportunity for the antigen determination of previously untypable isolates; while on the other hand it also leads to cases of incongruent results between the phenotypic and genetic serovar designation, specifically in regards to some important monophasic serovars. Overall these isolates were designated as genotypic matches, as the result provided by SISTR was genetically correct, but was not a match to phenotypic expression.

4.1.1.2.1 Understanding the Rough-O Isolates

Salmonella isolates with a rough-O or untypeable designation are occasionally found and reported to the national surveillance system. Previous validation studies of alternative serotyping methodologies have rarely considered these rough isolates, as they infrequently cause disease (27), but the identification of unexpressed antigens has generally been discussed as advantageous (51, 76, 77, 113). In all, we tested 25 rough-O isolates in our initial validation panel, of which 20 were from ssp I and the rest from the various non-ssp I *Salmonella* of clinical relevance. These isolates also displayed various levels of reporting for H-antigens, including some with a complete flagellar antigen designation to others that were considered completely untypable. SISTR provided some level of antigen determination for all 25 of the untypable isolates tested, including 20 which returned with a full O-antigen prediction.

Roughness in serotyping is the result of the loss of the sugar moiety on the LPS antigen and there is evidence that the ‘roughness’ reported by traditional serotyping is not the result of a single genetic loss or change, but instead the result of various mutations, frameshifts, and full

gene deletions within the *rfb* region (114). Sometimes the smoothness of an isolate can be recovered following additional steps in traditional serotyping (44), such as repeated passages through 0.38% Craigie tubes (data not shown), indicating that roughness may even be the result of inefficient polymerase activity (44). In the case of the galactose initiated serogroups (serogroups A, B, C2-C3, D1, D2, D3, and E) the functional loss of the *wbaP* gene is indicative of complete roughness, as this gene is responsible for adding the initial galactose sugar to the start of the O-antigen polysaccharide. Without the initial galactose unit, no additional sugar units can be added and the O-antigen component of the LPS is not produced (44, 83).

A total of 14 of the rough-O isolates in this study had an O-antigen prediction consistent with the galactose initiated serogroups, and of these 14, three had a completely non-functional *wbaP* gene. These isolates carried a mutation that lead to a premature stop codon within this gene. A further six isolates displayed one or more point mutations within the *wbaP* gene leading to amino acid changes and an additional isolate had frameshifting deletion of 67 bp. The mutations and deletions seen in these seven isolates were not seen amongst the panel of smooth galactose initiated serovars tested as part of the study, indicating these could potentially be responsible for the rough-O phenotype recorded; however, further work would be required to determine if these changes actually abrogate the functionality of the *wbaP* gene, or if some other mutation within the *rfb* operon was leading to this phenotype.

These findings are in line with the idea that roughness is not the result of a single genetic change, but that multiple possible changes can be responsible for this phenotype (44). As well no phylogenetic signal was detected amongst the group of rough isolates, instead the majority of these isolates strongly clustered with smooth isolates, indicating that these previously untypable isolates are not their own unique lineage, but are instead closely related to their smooth

counterparts, and have just lost the ability to express the O-antigen. The serovar determination of these rough-O and other untypable isolate indicates one major advantage of *in silico* serotyping, specifically the ability to classify a group of isolates deemed untypable by phenotypic methods.

4.1.1.2.2 Understanding the Phenotypically Monophasic Isolates

The other major group of incompatible isolates noted in the initial validation were monophasic variants of common serovars. This was specifically seen in what has traditionally been considered the monophasic Paratyphi B var. Java and the monophasic Typhimurium serovars that are recorded in NESP as serovars 4,[5],12:b:- and 4,[5],12:i:-, respectively. Both these monophasic serovars are important in the Canadian clinical setting and are responsible for many hundreds of cases of illness each year (27).

The 4,[5],12:i:- serovar was first reported in Spain in 1997 and has been found around the world (115). Despite its recent emergence, this serovar has been extensively discussed in the literature (100, 115-117). While multiple distinct mutations have been noted as being responsible for the deletion of the *fljB* locus, two main deletion patterns have been described as the Spanish and United States deletion genotypes, due to their predominance from those geographical regions (115). Additionally the 4,[5],12:i:- serovar has been shown to be largely monophyletic through 7-gene MLST analysis and is considered to be closely related to the diphasic serovar Typhimurium (76, 115). Meanwhile, the 4,[5],12:b:- serovar appears to have emerged even more recently, and although it is responsible for non-trivial amounts of disease in Canada and elsewhere (27, 101), it has not been as widely discussed in the literature. What has been noted is the polyphyletic nature of this serovar and its relation to more than just the diaphasic serovar, Paratyphi B var. Java. In fact, some 4,[5],12:b:- isolates have been shown to be related to lineages of serovars Abony (4,[5],12:b:e,n,x) and Mygdal (4,12:z91:-); however, the majority of these isolates are related to

Paratyphi B var. Java lineages (76, 101).

Interestingly, for both of these monophasic serovars, there has been discussion within the literature on a select group of phenotypically monophasic isolates that genetically still possess the *fljB* gene for the 1,2 antigen. These isolates can be considered phenotypically monophasic Typhimurium isolates that possess the *fljB* gene (PMT *fljB*+) and the phenotypically monophasic Paratyphi B var. Java isolates that possess the *fljB* gene (PMPBJ *fljB*+).

For the PMT *fljB*+ isolates multiple *IS*26 insertions into the promoter region of *fljB* have been detected that have abrogated its expression (100). Of the twenty 4,[5],12:i:- isolates tested, one was found to be PMT *fljB*+. This isolate displayed a truncated *hin* gene (Table 3) which would explain the non-expressed *fljB* gene; however, the detection of this truncated *hin* occurred at the end of an assembled contig so it is possible this is an artifact of assembly. Confirmation of the truncated *hin* in this isolate could be completed through a PCR of the region and Sanger sequencing or the development of a closed genome for this isolate, something which could be pursued in the future.

Meanwhile, the PMPBJ *fljb*+ have shown multiple various deletions within the flagellar phase variation operon, ranging from the deletion of the *hin* gene to a single bp deletion in a non-coding region downstream of *hin* (101). Of the 20 4,[5],12:b:- isolates tested, six were found to be PMPBJ *fljB*+. Five of these isolates showed a full gene deletion of the *hin* gene, while one displayed the same single bp deletion in the non-coding region downstream of *hin* that has previously been linked to the loss of *fljB* expression in PMPBJ *fljB*+ isolates (Table 3) (101).

Together these phenotypically monophasic *fljB*+ isolates provide further evidence for the need to recognize the incompatibilities between serotyping in the phenotypic and genomic eras. The identification of the multitude of mutations, deletions, and insertion events that explain the

phenotypically monophasic *fljB*+ isolates (100, 101) and the identification of other phenotypic monophasic *fljB*+ isolates that lack an explanation (101) really highlight how little phenotypic expression is understood. Fully understanding all the intricacies of the phenotypic expression of antigenic markers is beyond the current scope of biological knowledge, and would require detailed studies on the effects of various mutations across a multitude of genes to determine their effect (69).

It is important to note that the incompatibilities seen with the rough-O or the phenotypically monophasic *fljB*+ isolates are not the results of incorrect answers by either *in silico* serotyping or phenotypic serotyping, but instead reflect the different ways in which these methods approach the serotype question. *In silico* methods, such as SISTR, consider the carriage of the antigenic genes to define the serovar, while traditional serotyping techniques are solely concerned with the phenotypic expression of the antigens. Recognizing incompatible results between these methods represents a paradigm shift in conceptualizing serotypes, and understanding this shift is important as surveillance programs move towards using WGS to provide traditional subtyping results. It is therefore recommended that the results from these isolates be considered genotypic matches, thereby properly reflecting the differences in the testing methods, but recognizing the truth provided by both methods.

4.1.1.3 Expansion and Curation of the cgMLST Database Improves Results from SISTR

The expansion and curation of the SISTR cgMLST database, from which serovar calls are made, would address numerous issues that were identified in the initial validation of this platform. This included large-scale issues such as differentiating the subspecies of *Salmonella* and some important serovar variants to more small-scale issues that were uncovered such as the proper characterization of unique phylogenomic lineages of some serovars.

Ten percent of the isolates tested displayed poor cgMLST clustering results, which occurred when an isolate matched to a genome within the SISTR cgMLST database on less than 85% of the 330 alleles. The majority of these isolates were from the rare and unusual serovars tested as part of the panel; however, there were a few isolates from common serovars, such as serovar Oranienburg. This serovar is a known polyphyletic serovar with multiple highly distinct lineages circulating around the globe (76), and these isolates with poor cgMLST results indicate that unique lineages of Oranienburg circulating in Canada were not captured in the initial cgMLST database. Expanding the cgMLST database with representatives from these and other under-represented lineages would further enhance SISTR serovar prediction, and strengthen the results provided by a crucial part of the platform.

4.1.1.3.1 cgMLST Can Differentiate Non-subspecies I Isolates

Past investigations into non-traditional serotyping techniques have either not, or very briefly touched on the serotyping of non-subspecies I isolates (51, 77, 91). While the overwhelming majority of human *Salmonella* infections are caused by subspecies I isolates, non-subspecies I isolates can still lead to human disease (27). Most non-subspecies I *Salmonella* are considered environmental isolates and are mainly linked to cold-blooded animals (118, 119). However, for a national reference laboratory with surveillance functions, proper identification of these and other rare serovars remains a crucial part of their mandate.

The differentiation of *Salmonella* species and subspecies was not initially included within the SISTR platform. This lead to the improper categorization of many of the non-subspecies I isolates by SISTR. However, the 330 cgMLST scheme was able to separate out the two *Salmonella* species and the five *S. enterica* subspecies into well defined branches (Figure 4). Previous research on the genomic epidemiology of the genus *Salmonella* has noted the long

evolutionary history of this genus (118) and the strong linkage of the taxonomic groups to deep phylogenetic branches (118-121). It is therefore not surprising that the SISTR cgMLST scheme was able to differentiate the taxonomic groups of the genus *Salmonella* from each other, and incorporating these data into the SISTR serovar prediction algorithm would provide stronger results for non-subspecies I serovars.

4.1.1.3.2 cgMLST Can Differentiate Serovar Variants of 4,[5],12:b:1,2

Multiple *Salmonella* serovars possess unique serovar variants, in which a single antigenic formula can be further subdivided based on biochemical tests (13). The initial SISTR platform was unable to differentiate beyond the serovar level and provided no further information on important serovar variants (41). The antigenic formula 4,[5],12:b:1,2 possesses two variants, named Paratyphi B and Paratyphi B variant Java, which are differentiated by their ability to ferment the carbohydrate *d*-tartrate. Isolates that cannot ferment *d*-tartrate (dT-), have traditionally been associated with a more severe typhoid-like systemic disease, while those that can (dT+) are associated with a milder gastrointestinal syndrome (102, 122, 123). The lead acetate test is considered to be the gold standard biochemical test for assessing the fermentation of *d*-tartrate; however, this test is highly problematic as it is difficult to perform, requires a long incubation time (up to 14 days), and can give variable results (102, 122). Altering the incubation conditions of the lead acetate test has improved its reliability and false negative rate, but its long incubation time and complex procedures still present a barrier (102).

Alternative methodologies for assessing an organism's ability to ferment *d*-tartrate have been developed in attempts to address these issues. Most notably a PCR assay was developed to discriminate between fermenters and non-fermenters. This assay was developed around the functionality of a gene for a putative cation transporter in *Salmonella*, which resides at a locus

with homology to *l*-tartrate dehydratase genes in *E. coli*. It was found that the loss of function in this gene, through a G to A mutation at the start codon, resulted in a loss of the ability to ferment *d*-tartrate. Unlike the lead acetate test, the PCR protocol was able to identify all *Salmonella* non-fermenters reliably in a matter of hours versus the days required to perform the lead acetate test (102).

Multiple studies have also indicated the polyphyletic nature of this antigenic formula which is believed to be due to recombination events amongst flagellar antigen genes from various lineages with the 4,[5],12 O-antigen (123). Interestingly all forms of phylogenetic inference, from multilocus enzyme electrophoresis to WGS, have shown that the systemic variants of this antigenic formula form a discreet cluster that is separate from the enteric variants (76, 122, 123). This is greatly in line with the findings from the SISTR cgMLST tree for this antigenic formula (Figure 2) when the isolates were curated on their *d*-tartrate fermentation status via an *in silico* PCR result.

The most convincing evidence for a phylogenetic signal of the increased invasive potential of the dT- group comes from a WGS study of its antigenic formula (123). This study confirmed that isolates with the dT- SNP formed a unique phylogroup. Metadata on these and other isolates in the study uncovered multiple instances of invasive infections from both dT- and dT+ isolates with this antigenic formula; however, only the dT- phylogroup displayed a strong association with invasive infections. All other phylogroups of dT+ isolates were not strongly associated with invasive infections, and instead were mostly collected from enteric infections (123).

Other attempts to address the issue of the systemic and enteric variants of this antigenic formula have focused on attempting to identify the virulence genes associated with the systemic

variants. This ultimately presents a more direct test of the important characteristic that sets these variants apart from each other versus using a proxy test based on fermentation status. Examinations revealed that the isolates responsible for systemic disease possessed the virulence effector protein gene *sopE1* but lacked the virulence effector protein gene *avrA*, while isolates responsible for enteric disease possessed a different combination of these genes, most often lacking the *sopE1* gene and possessing the *avrA* gene (122). However, the Connor *et al* (2016) study did not find this to be true amongst their larger data set (123). It should be noted that we also did not find this to be true amongst our dataset. While all our dT- isolates possessed *sopE1* and lacked *avrA,* we also detected this pattern in four other isolates that were dT+ and clustered well away from the dT- isolates. While it is possible this combination of virulence effector proteins could signal enhanced invasive potential, it is not a strong signal for the differentiation of Paratyphi B from Paratyphi B var. Java, and for now, *d*-tartrate fermentation remains a necessary step for clinical assessment pending identification of the genotypic characteristics that underlie invasiveness.

Ultimately the identification and proper categorization of the systemic and enteric varieties of this antigenic formula remains important to public health officials for assessing risk and carrying out surveillance activities. Contact tracing and follow-up is often required for infections of *Salmonella* Paratyphi B to limit its spread but is not undertaken for *Salmonella* Paratyphi B var. Java (123). In Canada, the antigenic formula 4,[5],12:b:1,2 is of increased clinical importance due to its widespread nature, and the majority of disease caused by this antigenic formula has been the result of the Paratyphi B var. Java variant (27). While SISTR was unable to initially identify or delineate between these two variants we found multiple instances of improperly coded and classified isolates within the SISTR cgMLST database. These

improperly classified isolates within the cgMLST database were skewing the ability of SISTR to differentiate these isolates using the cgMLST platform. Of the 40 isolates with this antigenic formula in the database only 19 were properly labeled based on their *d*-tartrate fermentation status, determined through an *in silico* PCR test. Curation of these isolates within the cgMLST database allows for the use of the cgMLST results to infer the serovar variant status of this antigenic formula (Figure 2), providing key information needed and required by public health laboratories performing serotyping tests.

4.1.1.3.3 SISTR Can Detect Serovar 4,[5],12:b:-

A further complicating aspect of the antigenic formula 4,[5],12:b:1,2 is the presence of monophasic isolates of Paratyphi B var. Java, or isolates that only possess the H1 flagellar antigen, named serovar 4,[5].12:b:- (76, 124). This serovar designation was not initially included as a possible serovar within SISTR, and only two serovars, Schleissheim (antigenic formula 4,12,27:b:-) and ssp II 1,4,[5],12,[27]:b:[e,n,x], were considered possible options for the antigenic formula of B:b:-. Due to this developmental issue nine publically available genomes that displayed this antigenic formula by SISTR were erroneously curated to serovar Schleissheim within the SISTR cgMLST database. Both the erroneous curation and the developmental oversight resulted in incorrect predictions for the 4,[5],12:b:- isolates tested in the initial validation of SISTR.

Serovar Schleissheim is a rare serovar and can often be confused with 4,[5],12:b:- serovar (42). Literature searches have only shown three human cases of disease linked to this serovar, all originating in Asian nations (103-105), and a further 12 reports of environmental isolation (103, 125-127). Serovar 4,[5],12:b:- on the other hand is a common cause of human disease in Canada and is responsible for more disease than its diphasic variants in Canada (27). Serovar 4,[5],12:b:-

is also very diverse with multiple emergences and lineages in circulation (76, 101, 123). The analysis of STs and sequencing of the flagellar loci have shown that the 4,[5],12:b:- isolates are not just monophasic variants of Paratyphi B var. Java, but also can be monophasic variants of serovar Abony (1,4,[5],12,[27]:b:e,n,x) and even examples of flagellar antigen switches of named monophasics such as serovar Mygdal (4,12:z91:-) (101).

The population structure of the 4,[5],12:b:- isolates in this study also show evidence of this diversity (Figure 3). It is interesting to note that two publically available genomes with the reported serovar of Schleissheim cluster within the 4,[5],12:b:- group. It is likely that these two isolates are not actually serovar Schleissheim but in fact represent improperly reported 4,[5],12:b:- isolates. These isolates cluster with other 4,[5],12:b:- isolates in the SISTR cgMLST scheme, and their reported 7-gene MLST STs are linked to common 4,[5],12:b:- STs which are vastly different from any reported ST of serovar Schleissheim (76).

Following the initial validation of SISTR, developers incorporated serovar 4,[5],12:b:- as an option in the serovar prediction algorithm, and the nine public genomes erroneously curated in the SISTR cgMLST database to serovar Schleissheim were properly curated to serovar 4,[5],12:b:-. Lastly, the two isolates with a reported serovar of Schleissheim were removed from the cgMLST database until their reported serovar can be confirmed.

4.1.1.4 The Continued Curation of the Flagellar Antigen Databases is Needed

During the initial investigation into SISTR a total of five novel gene variants for flagellar antigens and one incorrectly labeled antigenic gene sequence was found. These novel gene variants were from a variety of H-antigens from uncommon serovars and lead to incorrect results in the initial SISTR validation. While this was a minor problem overall, these incorrect results point to the need for continued curation of the antigenic gene determinant databases, as well as

the importance of running SISTR and traditional serotyping in parallel for rare and unusual serotypes to ensure predictions from SISTR are accurate going forward.

4.1.2 Comparison of *in silico* Serotyping Prediction Methods

In this study three platforms, SISTR, SeqSero, and 7-gene MLST, were all assessed on their ability to provide *in silico Salmonella* serovar predictions. Successful results were reported for 94.8%, 88.2%, and 88.3% of the 813 isolates tested using SISTR, SeqSero, and 7-gene MLST, respectively (Table 5). While all platforms displayed a high overall success rate the breakdown of the successful results into more informative categories provides a better picture of the results from the platforms.

4.1.2.1 Partial Matches Require Further Analysis to Provide a Full Serovar Call

All platforms provided partial matches for some isolates. The percentage of isolates that returned with partial matches ranging from a low of 1.1% using SISTR to a high of 30% using SeqSero (Table 5). The majority of partial matches recorded by SeqSero was specifically due to ambiguous results and occurred when multiple serovars had the same or similar antigenic formula. This was seen in the case of Carrau (6,14,[24]:y:1,7) vs. Madelia (1,6,14,25:y:1,7) serovars that differ in the presence of minor O-antigenic factors, or in cases such as Javiana (1,9,12:l,z28:1,5) vs. the ssp II serovar 9,12:l,z28:1,5, which differ in their sub-speciation. There is some evidence that minor O-antigenic factors, such as O:6, is variable expressed, implying serovars such as Hadar (6,8:z10:e,n,x) or Istanbul (8:z10:e,n,x) are actually one and the same (128). Therefore, the ambiguous results provided by SeqSero for isolates that differ on the O:6 minor antigenic factor could represent full matches. An updated WKL scheme could potentially resolve these issues and reduce the number of inconclusive matches recorded by SeqSero. Other partial matches provided by SeqSero were the result of a lack of subspecies designation, a lack of

serovar variant determination, the exclusion of monophasic isolate designations, and missing information from O-antigen calls (due to the issue of size selected data).

Meanwhile, the ability of SISTR to resolve ambiguities in the antigenic calls of serovars, determine serovar variant status, and provide subspecies identification is only as good as the data that is included within its cgMLST database. In the initial validation of the SISTR platform instances of poor cgMLST clustering, improperly curated public genomes, and a lack of reporting crucial information was found. It is therefore important that the SISTR cgMLST database is continually updated and curated to ensure it is providing the best possible information. The few cases of partial matches provided by SISTR was the result of an ambiguous call, due to an inconclusive phylogenomic context. In all but one of these cases the reported serovar was incredibly rare and/or unusual, and therefore had no representatives within the cgMLST database.

Lastly, all the partial matches provided by 7-gene MLST were the result of no linkage between the 7-gene ST determined and any serovar within the public database. The majority of these cases were from rare and/or unusual serovars, especially the non-subspecies I serovars that were tested. This finding is unsurprising due to the limited coverage within the 7-gene MLST database for these isolates and was an issue reported in the initial validation of serotyping through 7-gene MLST (77). These results represent a potential area of improvement with 7-gene MLST serovar prediction, as databases that link 7-gene ST profiles and serovars could continually be expanded to include information on more rare and unusual serovars.

Ultimately, isolates that returned with a partial match would require further analysis either by using the traditional serotyping techniques, other biochemical tests, or further genomic analysis to provide a full serovar call. For some users, partial results may be of little concern as

the designation provided by the platform in these cases does not provide any incorrect information, and other than with the partial matches from 7-gene MLST data, where no information was reported, the results provided by the *in silico* platforms are informative and can be used in the identification of the correct serovar. However, for other users who are interested in a full answer, partial matches could present a complication and may require the maintenance of the technical and logistical capacity to carry out traditional serotyping or require submitting the isolate to a reference laboratory where the generation of results may be delayed.

4.1.2.2 Genotypic Matches Require the Adoption of a New Paradigm for Defining Serovars

A total of 33 isolates tested, regardless of the platform utilized, provided a genotypic match (Table 5). For these isolates, the genetic serovar determination is incompatible with their phenotypic serovar designation, due to the carriage of unexpressed antigenic markers. As more fully discussed in section 4.1.1.2 these isolates either had a rough-O or untypable serovar designation, or were phenotypically monophasic, but genetically diphasic. For the 26 isolates with a rough-O or untypable designation by traditional techniques, complete serovar calls were generated for 96%, 54%, and 85% of the isolates tested by SISTR, SeqSero, and 7-gene MLST, respectively. While partial results were generated for all these isolates in SISTR and SeqSero.

As previously mentioned, the generation of results for the Rough-O and untypable isolates has been considered a positive application of alternative serotyping techniques, including *in silico* methods (51, 76, 77, 113). Therefore, the generation of complete or partial serovar call would be considered a large advantage of using *in silico* methods for serovar determination. The consistency of the results between the platforms and with any reported H-antigens is also a positive result. However, the inconsistency between one isolate's 7-gene MLST prediction and its reported H-antigens could be considered an incorrect result.

The determination that generating a serovar call for rough-O or untypable isolates is an advantage of genetic serotyping also requires us to recognize that the same logic must be applied to any other unexpressed antigenic determinant. Therefore, phenotypically monophasic isolates that carry a *fljB* gene should be called as the diphasic serovar name from which they evolved, and results from these isolates should also be seen as advantageous. It was for these reasons that isolates from these categories were considered genotypic matches and were considered to have provided successful results by all of the *in silico* serotyping methods used in this study.

Overall, the genotypic matches uncovered in our validation point towards one of the major implications of using any WGS-based analysis platforms for serotyping, specifically, that a genomic test is not a full equivalent to a phenotypic test in all cases. Using genomics to answer the serovar question requires us to adopt a new paradigm where the carriage of genes for O- and H-antigens determines the serovar, not the expression of these genes. As we move towards WGS as a replacement for serotyping it is important that downstream consumers of this information accommodate for the fact that there may be minor changes in the prevalence of key serovars due to isolates possessing antigenic genes that are not expressed. It is important for users of this information to note that the incompatibilities between the *in silico* and traditional serotyping results for these isolates are not indicative of incorrect answers, but the result of different ways of framing the serotyping question. As surveillance records are important documents for a multitude of users the identification of how the results were generated, version numbers of platforms used, and any limitations of the methods are crucial to ensuring a consistent basis of understanding.

4.1.2.3 Incorrect Results Highlight Areas for Further Development

All analysis platforms reported the presence of some incorrect results amongst the 813

isolates tested, ranging from a low of 5.2% of isolates tested using SISTR to just over 11% of isolates tested using SeqSero and 7-gene MLST (Table 5). The non-target serovars from panel three provided the most number of incorrect results for any serovar grouping from all platforms (Supplementary Table B). Interestingly, incorrect results were recorded for a group of thirteen isolates, all from panel three, regardless of the platform utilized. For twelve of these isolates the same incorrect call was made across all three platforms, potentially calling into question the reported serovar for these isolates. However, the confirmation of the reported serovars for these isolates was not pursued by the Guelph team for these isolates.

For both SeqSero and SISTR some incorrect results were due to incorrect calling of various antigenic determinants, especially in regards to closely related serovars, such as those that differ on the basis of flagellar antigens of the g-complex. This is a limitation of both traditional serotyping (48) and molecular-based serotyping techniques (50) and is related to the high sequence and amino acid similarity amongst some flagellar antigens of the g-complex (81). As well, isolates with novel gene variants that were found during the initial validation of SISTR all returned with a wrong call in SeqSero. This points to the fact that the gene databases are only as strong as the data stored in them, and that SeqSero has the potential to improve their accuracy through the addition and curation of these novel flagellar genes.

The number of incorrect calls for SISTR and SeqSero was also impacted by the poor sequencing coverage over the *rfb* region due to the size selection of the genomic libraries. Unfortunately, all the data generated for this validation was impacted by this process and lead to the majority of the wrong calls from SeqSero and some wrong calls from SISTR. As previously discussed, the low sequencing coverage over the *rfb* region was producing assemblies where there was a split over the *rfb* region, often at the highly variable *wzx* and *wzy* genes. Since there

is a high degree of similarity in the *rfb* region outside the *wzx* and *wzy* genes for many O-antigens (83), and SeqSero makes an O-antigen call using data from the entire *rfb* region (42), many isolates returned with wrong O-antigen calls. Meanwhile, SISTR was only looking at the *wzx* and *wzy* genes when making its O-antigen call (41), so when this information was missing no call could be made. Strong cgMLST serovar predictions congruent with the flagellar antigens would often lead to a correct serovar prediction in these cases, thereby reducing the negative effects of this sequencing issue on SISTR. However, this sequencing issue did lead to some incorrect calls in SISTR, specifically for isolates from the D1 serogroup that displayed the poor sequencing coverage over the *rfb* region. For these isolates, SISTR provided an O-antigen serogroup prediction for the B serogroup. This was due to the carriage of the serogroup B *wzy* gene at a separate locus within the genome, a phenomenon unique to the D1 serogroup (83). Future implementation of *in silico* methods for *Salmonella* serotyping using either SISTR and especially SeqSero should ensure this size selection step is skipped in the library preparation stage to prevent any bias in the sequencing library.

Lastly, incorrect results from both SISTR and 7-gene MLST analysis were also attributed to a lack of sufficient representative isolates from a specific serovar/and or phylogenic lineage, throwing off the prediction. In many instances, this points to areas of improvement for both these platforms through the curation and expansion of their databases to accommodate more rare and unusual serovars and unique phylogenic lineages. However, this issue is not just attributable to the rare and unusual serovars. 7-gene MLST analysis provided an incorrect prediction for all 4,[5],12:i:- isolates. Since this method determines a serovar based off of the dominant serovar within the database at a specific ST/eBG (76, 77), and all isolates of serovar Typhimurium and 4,[5],12:i:- belong to the same eBG (76), an incorrect call for 4,[5],12:i:- isolates was recorded.

In SISTR, a similar issue was noted for some 4,[5],12:i:- and 4,[5],12:b:- isolates which clustered closely with Typhimurium and Paratyphi B var. Java representatives, respectively. In this case, the antigenic call of the monophasic variant was overridden by cgMLST clustering.

4.1.2.4 Choice of Serovar Prediction Software Depends on Users Need

SISTR significantly outperformed both SeqSero and 7-gene MLST analysis in providing successful results for the *in silico* serotyping of isolates in this study. This was also the first time all three platforms were tested on the same dataset of isolates. Previously, all platforms had been validated by their developers and reported accuracies of 94.6% for SISTR (41), 92.6% for SeqSero (42), and 96% for 7-gene MLST (77). Although we found significantly lower levels of accuracy reported by the accuracy rates from SeqSero and 7-gene MLST validations, all platforms performed significantly better than the accuracy of traditional serotyping, which was reported as a global average of 82% (48). Lower levels of accuracy in this study for the SeqSero platform in comparison to its initial validation could potentially be explained by the sequencing issue at the *rfb* region. While the lower level of accuracy for 7-gene MLST in this study in comparison to PHE's validation could be explained by their use of an internal and much larger database linking ST to serovars. Overall the significant improvement of all platforms from the reported global accuracy of traditional serotyping shows the overall suitability of replacing phenotypic serotyping with an *in silico* analysis tool.

Sensitivity and specificity of the platforms for serovars Enteritidis and Typhimurium were also assessed. Sensitivity ranged from a low of 81.0% for serovar Enteritidis using the SeqSero to a high of 100% for Typhimurium using both SISTR and 7-gene MLST platforms, with the rest of the values falling between 95.2% and 97.4%. While specificity ranged from a low of 97.2% for Typhimurium using 7-gene MLST to a high of 100% for the same serovar

using SeqSero, with the rest of the values falling between 99.2% to 99.9% (Table 6). In all, these data indicate that all platforms show a high degree of accuracy for detecting the two most commonly reported serovars to public health laboratories worldwide (47). The low sensitivity (81.0%) for serovar Enteritidis using SeqSero was attributed to the *rfb* sequencing issue which lead to incorrect O-antigen calls for some Enteritidis isolates. The low specificity (97.2%) for serovar Typhimurium using the 7-gene MLST platform could be attributed to the grouping of serovars Typhimurium and 4,[5],12:i:- into a single category by this method (76).

The use of an *in silico* serotyping tool would be inappropriate for the investigation into the expression of antigenic factors, but all platforms indicated their suitability in maintaining the historical records and communication structures on which *Salmonella* surveillance is based due to their high level of accuracy, especially in regards to the clinically important serovars. However, all *in silico* platforms are limited in their ability to define novel serovars and/or antigenic genes, as their databases are only as strong as what is located in them. The further refinement of all platforms is possible, and the future parallel analysis of phenotypic and *in silico* methods, especially for rare and unusual serovars would be best suited to uncover any remaining gaps. Overall, the choice of a platform should be made based on the platforms relative strengths and weaknesses and the laboratories intended purpose.

SeqSero provides a genomic understanding of the individual antigens that make up a serovar (42), and represent the closest analogous situation to traditional serotyping. One potential advantage of SeqSero is the possibility of determining the serovar directly from the raw sequencing reads (42). While the raw read analysis through SeqSero was not assessed in this study, it may represent a positive for laboratories that lack the computational capacity to assemble genomes prior to serovar determination. While the 7-gene MLST analysis is not

analogous to traditional serotyping it allows for enhanced phylogenetic information to be generated that can be used to answer additional epidemiological questions. For example, this method allows for the differentiation of *Salmonella* serovar Newport, a polyphyletic serovar, into the three distinct lineages allowing for the potential to further classify polyphyletic serovars (76). As well, this method could be adapted for the analysis of results from raw read data (77), also reducing the computation capacity for laboratories. Meanwhile, SISTR not only allows users to gain an understanding of the underlying genetic carriage of the individual antigens, but also allows for the generation of enhanced phylogenetic information that can further classify the isolates (41). The combination of both phylogenetic and genomic determinations of a serovar by SISTR allows for significantly stronger results and a reduction in the number of inconclusive matches that would require further benchtop serotyping.

4.2 Examination of WGS for *Salmonella* Outbreak Investigation

4.2.1 Serovar Hartford

Serovar Hartford is rarely linked to human clinical disease both within Canada (27), and elsewhere (129). Previous recorded outbreaks related to this serovar are rare, and interestingly the only other previously published outbreaks involving *Salmonella* Hartford were also multi-serovar outbreaks (129) (https://www.cdc.gov/salmonella/2010/restaurant-chain-a-8-4-10.html); however, the relevance of this is not apparent.

Increases in the disease frequency of this serovar were detected by NESP within Canada in April/May 2014, signifying the start of the sprouted chia outbreak. By the end of the outbreak investigation, 22 clinical cases of salmonellosis and four food isolates were identified from serovar Hartford all of which displayed two highly related PFGE patterns (Unpublished internal report, PHAC, 2014). WGS-based retrospective analysis of this outbreak revealed a single highly

related cluster consisting of the 26 isolates identified during the outbreak investigation that was separated by a maximum of six hqSNVs or alleles. One additional non-outbreak isolate was also found within this cluster. This isolate had a matching PFGE pattern to the outbreak case definition but was excluded from the investigation as it was identified after the official end date of the investigation. The discovery of additional outbreak cases past the date the outbreak was declared over is not surprising due to the numerous products implicated and their long shelf lives (107). Had the outbreak investigation been extended this case would have been included in the investigation, due to its matching PFGE pattern. Although the WGS-based analysis did not change the outcome of the investigation for *Salmonella* Hartford, additional clusters of disease from this serovar were noted among the circulating non-outbreak isolates, signifying additional events were ongoing within this serovar that went undetected by PFGE-based analysis.

4.2.2 Serovar Newport

Salmonella Newport is responsible for about two percent of all reported cases of Salmonellosis each year within Canada (27). Past recorded outbreaks of Salmonellosis linked to serovar Newport are not uncommon and have been linked to a diverse set of food products, including produce, dairy, and meat (10, 130-132). *Salmonella* Newport is also a polyphyletic serovar, with three distinct lineages known to be circulating globally (67) separated by many thousands of hqSNVs or alleles (133). Lineage I is most commonly reported within human populations in Europe (67), and only one isolate in the retrospective analysis was found to be from lineage I, indicating a potential travel history for this case. Lineage II and III are much more commonly reported within North America (67), and their dominance amongst the isolates sequenced as part of this investigation is not surprising. Past WGS-based studies of the Newport II lineage has indicated distinct subclades within this lineage (133), all of which are apparent and

circulating within Canada. It has been suggested that the lineages of *Salmonella* Newport diverged early on in serotype evolution and are as divergent from each other as they are to other serotypes (133), all of which was apparent in our analysis.

The *Salmonella* Newport isolates linked to the sprouted chia outbreak were all from lineage III, and both SNVPhyl and wgMLST placed these isolates into two distinct clusters that were split by over 100 hqSNVs or alleles, respectively. The smaller of the two clusters contained just three food isolates, with no human matches, and these isolates were identical via SNVPhyl and differed on just one allele via wgMLST. The larger cluster contained all 22 human clinical case isolates, plus 19 food isolates. One additional non-outbreak isolate also clustered with these 41 isolates. Once again, this isolate had a matching PFGE pattern to the case definition but was identified after the outbreak investigation, signaling the extended nature of this outbreak.

The larger cluster was separated by seven hqSNVs or fifteen alleles using SNVPhyl and wgMLST based analysis respectively. It should be noted that the majority of isolates in this cluster were separated by a maximum of two hqSNVs or six alleles, and just a single isolate, Newp-065, was separated from the rest of the clade by the seven hqSNVs or the fifteen alleles. The large discrepancy between the number of hqSNVs and alleles separating these isolates can be explained by the carriage of plasmids in Newp-065 that were not found in the other outbreak isolates. Since wgMLST captures diversity both within the core and accessory genomes, any plasmid-based loci will lead to differences among a group of isolates based solely on their presence in some and absence in others. Examining the individual loci that were different between Newp-065 and the other outbreak isolates in BioNumerics showed that seven of the fifteen allelic differences found were from core genome based loci. The others were from previously defined *Salmonella*-based plasmids, such as pCVM22462(134), pSC138 (135), and

pSLT-BT (136), which have all been shown to carry various antibiotic resistance elements. Since SNVPhyl only examines the core genome, diversity across the accessory genome, such as plasmids, are not considered (86). Therefore, the increased number of alleles separating this isolate from the other outbreak isolates is not shocking.

Overall WGS-based analysis did not change the categorization of cases in the sprouted chia outbreak investigation for serovar Newport, but it did provide a higher level of investigative resolution. However, the WGS-based analysis did alter other PFGE-based clusters of *Salmonella* Newport found during this time. One isolate would have been ruled out of cluster PNC-Newp-3, while two additional isolates would have been ruled into cluster PNC-Newp-4. For cluster PNC-Newp-4, the two additional isolates that would have been ruled in via WGS-based analysis were both identified earlier; thereby indicating a case where WGS would have detected a cluster sooner. The detection of this cluster at an earlier time point would have resulted in subsequent follow-up for this group of isolates. As well, one of the two additional isolates that would have been ruled into cluster PNC-Newp-4, specifically isolate Newp-141, was the first isolate identified in a different PFGE-based cluster from Ontario. Unfortunately, the other three isolates from the Ontario cluster that were ruled into this cluster by PFGE analysis were not sequenced as part of this study, so their relationship to PNC-Newp-4 cannot be determined, but this finding highlights the potential that a much larger event was ongoing that was not detected. Lastly, eight additional clusters, across all Newport lineages, undetected by PFGE analysis, would have been identified through WGS-based analysis, providing further evidence of missed avenues for follow-up investigation by PulseNet Canada.

4.2.3 Serovar Oranienburg

Salmonella Oranienburg is also a polyphyletic serovar (76) that is responsible for a non-

trivial amount of human clinical disease each year within Canada, occasionally entering into the top ten most commonly reported serovars (27). Serovar Oranienburg has also been reported in multiple past outbreaks linked to a variety of food products, such as fruit and chocolate (33, 137, 138). The *Salmonella* Oranienburg isolates sequenced in this study were highly diverse, and multiple lineages were found that were separated by many thousands of hqSNVs or alleles; providing further evidence of the polyphyletic nature of this serovar (76). Interestingly, the seven isolates of serovar Oranienburg identified in the sprouted chia powder investigation were from two lineages. Five of these isolates were from the main Oranienburg lineage in Canada; however, only two of these five isolates formed a cluster that was separated by ten hqSNVs or alleles. Both of these isolates had a PFGE pattern match to a sprouted chia powder product isolate that was not collected in Canada or sequenced as part of this study.

The other three isolates identified in the investigation consisted of one food isolate and two clinical cases that were considered PFGE matches; however, WGS-based analysis separated these isolates by more than 140 hqSNVs or 125 alleles, indicates that these isolates are not related to each other. It is important to note that the two clinical cases could still be related to the sprouted chia outbreak, but just be from unique strains of *Salmonella* Oranienburg that were not sampled from any food product. Multiple different strains were found within the sprouted chia products during the outbreak investigation, including many from serovars or PFGE patterns that did not have any clinical case matches (107). In fact, based on the WGS-based analysis a total of seven different food isolate clusters were found from the four serovars investigated, five of which had clinical cases matches. It is, therefore, possible that sampling of the food products during the food investigation failed to capture the full complement of *Salmonella* isolates that were present in the numerous sprouted chia products, and these two human clinical cases from

serovar Oranienburg were from these unsampled strains.

It is also possible that the *Salmonella* Oranienburg isolates identified underwent massive diversification during infection. This phenomenon has been noted in previous outbreaks of *Salmonella* Oranienburg in veterinary hospitals, where multiple isolates displaying multiple PFGE patterns were collected from a single animal infected with *Salmonella* Oranienburg during the course of infection. It was proposed that diversification, through the loss or acquisition of mobile genetic elements, recombination, or point mutation, may have occured and could explain the numerous collected PFGE patterns. However, co-infection by multiple strains could not be ruled out (139). Therefore, it is a possibility that these unrelated isolates in the sprouted chia outbreak could represent a case of massive and rapid diversification; however, the most simplistic explanation of the WGS-based data would be that these two human clinical cases were unrelated to sprouted chia powder outbreak.

The final two *Salmonella* Oranienburg isolates identified in the sprouted chia powder outbreak investigation were collected from food samples, and displayed unique PFGE patterns that were unrelated to any human clinical cases. These two isolates clustered well away (separated by many thousands of hqSNVs or alleles) from any other *Salmonella* Oranienburg isolates collected during this time period; however, both these isolates were identical to each other by both WGS-based analyses.

The one other PFGE identified *Salmonella* Oranienburg cluster from this time period was fully resolved by both WGS-based analysis methods, as all three PNC-Orani-1 isolates were found to be highly distinct from each other. WGS was also able to identify six additional clusters of two or more isolates that were separated by less than ten hqSNVs or alleles. Two of these clusters contained isolates of the most common serovar Oranienburg PFGE pattern combination

in this study, OraniXAI.0001/OraniBNI.0019. This indicates the great potential of WGS-based method to provide increased resolution and clustering to common PFGE pattern types, something which has been reported for other organisms (84) and serovars (140, 141).

4.2.4 Serovar Saintpaul

Like serovar Oranienburg, serovar Saintpaul is considered polyphyletic (76) and multiple lineages were noted to be responsible for disease in Canada during the outbreak period. In Canada, serovar Saintpaul is also responsible for a non-trivial amount of human clinical disease (27), and past outbreaks of Salmonellosis from this serovar have been noted and linked to various produce items (142-144).

Like serovar Hartford, increases in by *Salmonella* Saintpaul infections were noted in April/May by NESP; however, the full extent of the sprouted chia powder outbreak did not become apparent until later due to the low case density (Unpublished internal report, PHAC, 2014). WGS-based analysis revealed two separate clusters of isolates related to the sprouted chia powder investigation from serovar Saintpaul. These two clusters were separated by more than 70 hqSNVs or alleles. The smaller of the two clusters contained five isolates that were identical by both WGS-based analysis methods. Of these five isolates, four were collected from sprouted chia powder products with PFGE pattern profiles that did not match any clinical cases. The additional clinical case in this cluster had a matching PFGE pattern, but once again was excluded from the outbreak due to its identification after the investigation was closed. This isolate provides further confirmation that this was an extended outbreak (107) and that individuals were continuing to get sick past the official end date of the investigation.

The larger of the two sprouted chia powder clusters contained 21 *Salmonella* Saintpaul isolates that were identified during the outbreak investigation, which consisted of fifteen clinical

case isolates and six food isolates. This cluster was separated by a maximum of two hqSNVs or alleles. Once again, two additional isolates, Saint-082 and Saint-084, that had a matching PFGE pattern to the outbreak case definition were found within this cluster, but these isolates were excluded from the investigation as they were identified after the outbreak investigation was closed. Interestingly, an additional two clinical isolates, Saint-091 and Saint-054, also found within this cluster were excluded from the initial outbreak investigation, based on variant *Bln*1 patterns, indicating that two clinical cases were missed by the PFGE-based investigation.

Two additional PFGE-based clusters of *Salmonella* Saintpaul were identified during this time period. One of these clusters, PNC-Sain-1, was resolved through WGS-based analysis. Of the fourteen isolates included in this cluster by PFGE-based analysis, only four isolates were found to be related. Specifically, these four isolates were evenly split into two separate clusters, one of which consisted of two additional *Salmonella* Saintpaul isolates that were excluded from previous PFGE-based cluster analysis due to their differing *Xbal*1 pattern or temporal distance from the other isolates. The other PFGE-based cluster from this time period was fully confirmed by WGS-based analysis. Lastly, two additional clusters of *Salmonella* Saintpaul were identified indicating the improved cluster detection of WGS-based analysis.

4.2.5 Implications of WGS-Based Analysis for Outbreak Detection and Investigation

4.2.5.1 Choice of Analysis Platform Depends on Users Need

Overall both WGS-based methods assessed in this study, SNVPhyl and wgMLST, provided highly similar results in the assessment of the sprouted chia outbreak and the relevant non-outbreak comparators. Between the platforms, the same number of hqSNVs or alleles were found to define the isolates within the outbreak clusters, except for serovar Newport. This was due to the carriage of multiple plasmids by a single outbreak isolate that could only be detected

by wgMLST as this method looks at differences both within the core and accessory genomes. As well, similar numbers of hqSNVs or alleles were found to separate these clusters from the temporally related non-outbreak case comparators (Table 11).

Table 11: The number of hqSNV and allele differences used to define the isolates within the chia outbreak clusters, and minimum differences between these clusters and non-outbreak isolates.

Serovar	Maximum Differences within the Outbreak Cluster		Differences Between Cluster and Non-Outbreak Case Comparators	
	hqSNVS	**Alleles**	**hqSNVs**	**Alleles**
Hartford	6	6	1,297	936
Newport	7	15	565	504
Oranienburg	10	10	166	152
Saintpaul	2	2	71	64

Going forward, the choice of a platform will be decided by the user's need as both platforms provided an increased resolution to both the outbreak investigation and to cluster detection, where past clusters were resolved and additional clusters were identified that were undetected by PFGE-analysis alone. While both methods provide many of the same benefits, they can still be compared across other metrics (Table 12).

Table 12: Comparison of SNVPhyl and wgMLST platforms for phylogenomic analysis of *S. enterica* genomes.

Metrics	wgMLST	SNVPhyl
Requires a Reference	No	Yes
Requires a Schema	Yes	No
Speed	Slow allele calling Fast comparisons	Slow SNV calling Slow comparisons
Phylogenetic Information Provided	High	Higher
Standardized Nomenclature	Yes	Possible
Developmental History	Since 2016	Since 2010

The SNVPhyl platform requires the use of a high-quality closely related reference genome to provide its analysis (86). The choice of a reference in any reference-based analysis method is incredibly important as the reference can influence the SNVs that are called in numerous ways (145, 146). In fact analyzing the same outbreak using different references can

influence the SNVs that are called, making comparisons on SNV differences between these analyses impossible (85). A reference that is missing important areas of variation between the isolates being examined will obviously be unable to call these informative SNVs, and the resulting tree will make a group of isolates appear more closely related than they actually are (145). At the same time, a vastly different reference can also produce an excess number of SNVs between the reference isolate and the isolates under study, skewing the tree that is produced and hampering the ability to make a visual inference (146). The SNVPhyl tree produced from all 368 isolates sequenced in this study (Figure 7), shows both these issues amongst the various serovar clades. The reference genome in this 368 isolate tree was a fully closed reference genome from lineage II of serovar Newport, and the choice of this reference which was vastly different from not only the other Newport lineages but also the other serovars, influencing both the topology and SNVs found within the tree. The diversity among the isolates within the serovar groupings is difficult to determine due to large differences between these groups and numerous informative SNVs are also missing within these groups. For example, only 52 hqSNVs separate the two chia outbreak clusters from serovar Newport, while the SNVPhyl tree produced on these isolates with a more closely-related reference found almost 100 hqSNVs between these two chia outbreak clusters. It has been previously proposed for *S. enterica* that reference genomes should at least be from within the same serovar as the isolates under study (146). However, as noted the SNV distances between the various clades of polyphyletic serovars can be just as great as those between serovars, so the analysis of polyphyletic serovars requires the production of SNVPhyl trees of the different lineages with references from within those lineages.

wgMLST does not require the use of a reference genome and therefore bypasses this issue. However, for this analysis, a wgMLST scheme was obviously required and the issue of

undefined variation can still occur if the variation exists in previously undefined loci (147). The addition of previously undefined loci to an already existing wgMLST scheme is not recommended as this will alter the standardized nomenclature system, removing a key feature of this type of analysis (85). Instead during the development and validation of the schema, its robustness across a variety of serovars should be examined, to ensure its universal applicability. The *Salmonella* wgMLST schema used in this study was developed from over 260 reference genomes and includes over twelve thousand loci that were meant to capture the diversity both within the core and accessory genome of *Salmonella* (90). On average 4,412 loci were found in each genome regardless of the serovar examined. However, over nine thousand different loci were found between the various serovars and over five thousand different loci were found within each of the lineages examined indicating that many loci are serovar or lineage specific. Overall the ability of the *Salmonella* wgMLST scheme to provide the similar level of analysis amongst the four serovars from this study indicates its usefulness in providing phylogenetic analysis across diverse serovars.

Both platforms are relatively slow and require computationally intensive processes to call the SNVs or alleles used in the comparison. However, once a wgMLST profile of an isolate has been generated in the calculation engine that information is stored within the database and the pairwise comparison of the isolate and any others can be generated within seconds using computationally efficient distance-based algorithms (73, 88), making this analysis extremely amenable to adding or removing isolates from a comparison. The SNVPhyl pipeline does not display the same features as the entire pipeline must be re-run when adding or removing isolates to the comparison. The SNVPhyl developers are currently examining various ways in which isolates could be added or removed from a tree without requiring the rerunning of the entire

pipeline (Personal Communication, Morag Graham, 2017). However, the SNVPhyl pipeline utilizes a more computationally intensive character-based tree building methodology (73, 86) in comparison to the distance-based methods of wgMLST (73). In fact, the character-based algorithmic methods used by SNVPhyl can also require additional local computing power to complete its processes (85). However, these character-based methods do provide a wealth of additional information in comparison to any distance-based methods. Hypotheses can be generated for each nucleotide site in the MSA, allowing for the tracing of individual changes across the nodes, something which cannot be done using the wgMLST based data (88).

wgMLST data is also much more amenable to the generation of a standardized nomenclature system, allowing for the efficient communication of sequencing information between laboratories and jurisdictions. The production of a standardized nomenclature system is a highly valuable feature of any MLST-based typing system (85, 148). With wgMLST large amounts of sequencing information can be collapsed down to a single number or a string of numbers, using a prescribed allele and strain nomenclature. The allele nomenclature defines the loci and the sequences of the alleles, while the strain nomenclature is a specified classification system based on the various alleles identified (85). On the other hand, since the SNVs included in any SNV-based typing systems are influenced by both the isolates included in the tree and the reference genome, this platform is much less amenable to the development of a standardized nomenclature system (148). However, one could imagine that a SNV-based system could share information on important or potential clusters across jurisdictions through the sharing of raw sequencing read data and the use of a preestablished set of reference genomes for individual serovars/lineages. Alternatively, PHE is using a SNV-address approach where isolates are clustered via a hierarchical approach of increasing levels of similarity (149), similar to the strain

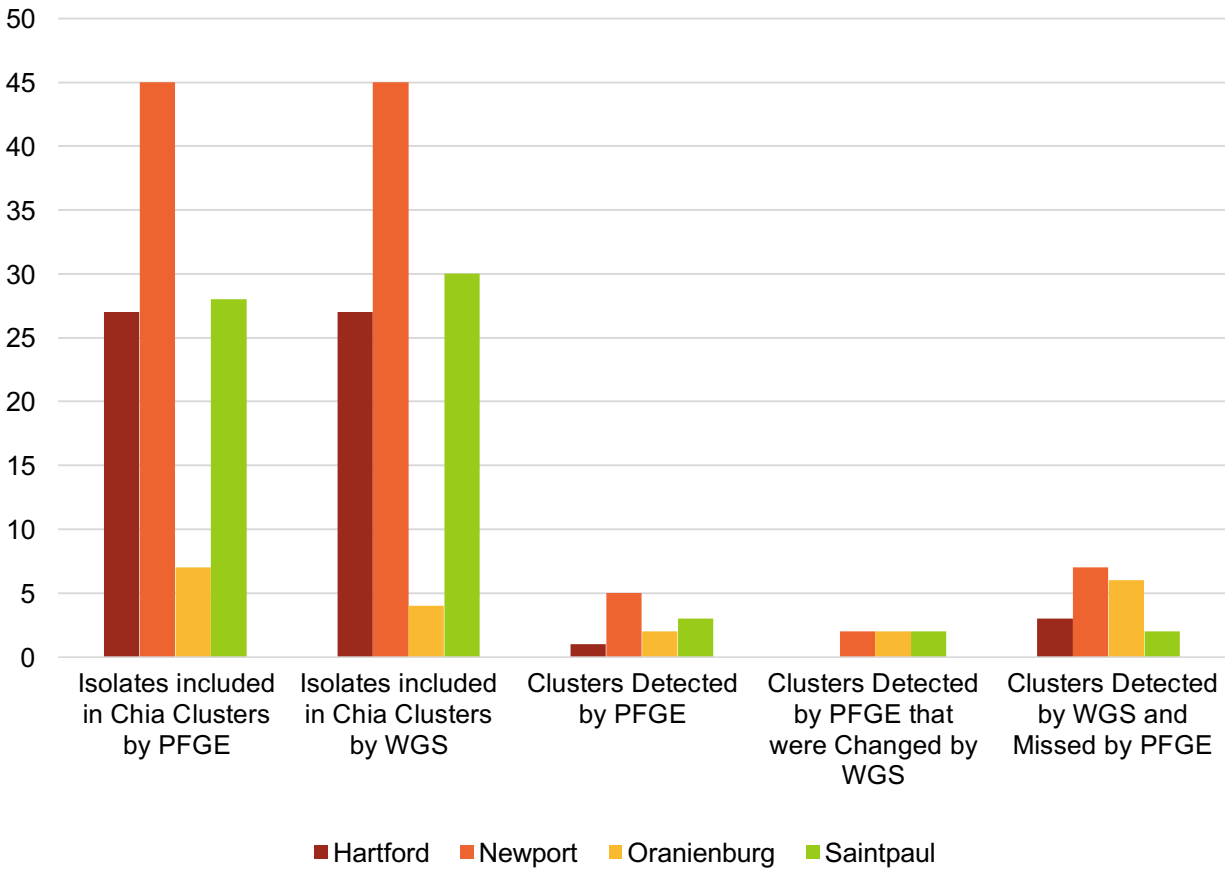
nomenclature system proposed by PulseNet International for wgMLST data (85).

One of the major strengths of the PulseNet International system is the ability to easily share important information across jurisdictions and borders, allowing for the identification of geographically-dispersed clusters (29, 85). This type of information sharing was crucial in the investigation into the sprouted chia powder investigation as a small cluster of unique PFGE patterns for *Salmonella* Newport was first reported in the United States that was linked to 'healthy eater' profile. The sharing of this information identified PFGE matches within Canada that had the same profile and shortly thereafter the entire investigation intensified (107). For this reason, and the success of the wgMLST pilot project for *Listeria* (84), PulseNet International is pursuing the development of standardized wgMLST schemes for priority pathogens (85, 150), creating stable allele and strain nomenclature systems that would allow for the fast and efficient communication of outbreak and cluster information between organizations. However, use of the SNVPhyl pipeline for a more in-depth analysis or confirmatory role within PulseNet Canada is feasible. Just like with the PFGE-based interpretation criteria (61), the availability of alternative subtyping data, potentially generated through alternative bioinformatic methods, to confirm results would be valuable and strengthen the weight of evidence in outbreak scenarios, as was the case in a retrospective outbreak analysis of listeriosis (89).

4.2.5.2 Improved Case Categorization using WGS Based Methods

Both WGS-based methods showed improvements over PFGE analysis in the case categorization of isolates in this study, regardless of serovar. Regarding the sprouted chia outbreak investigation, two new cases were identified from serovar Saintpaul, and two cases from serovar Oranienburg were excluded from the outbreak (Figure 17). Interestingly, the epidemiological investigation revealed two confirmed outbreak cases which reported travel to

Figure 17: The impact of WGS data analysis on the retrospective examination of four *Salmonella* serovars linked to an outbreak of sprouted chia seed powder in 2014.


50
45
40
35
30
25
20
15
10
5
0
Isolates included in Chia Clusters by PFGE
Isolates included in Chia Clusters by WGS
Clusters Detected by PFGE
Clusters Detected by PFGE that were Changed by WGS
Clusters Detected by WGS and Missed by PFGE
Hartford
Newport
Oranienburg
Saintpaul

Mexico during their entire exposure period, and there was some initial concern these cases could have been non-outbreak related matches to the PFGE pattern combination (Unpublished internal report, PHAC, 2014). However, both cases were found among the outbreak clusters in the WGS-based analysis, and with the increased resolution that WGS data provides in determining pathogen relatedness, it can be confirmed that the right choice was made in ruling these isolates into the outbreak. In addition, case categorization was improved for other PFGE clusters identified by PulseNet Canada and multiple additional clusters were identified among all serovars that were missed by PFGE-based analysis (Figure 17).

The improved case categorization by WGS also expanded past clusters and potentially would have resulted in their more timely identification by PulseNet. This was seen with the PFGE-based cluster PNC-Newp-4 and the additional isolates that were identified using WGS-based analysis. As well, one of these two additional isolates was also the first isolate uploaded to a different PFGE-based cluster that was not sequenced in this study. There is a strong potential that some larger event was occurring that was undetected by PFGE-based analysis due to the differing *Xba*1 patterns found in these isolates.

With any surveillance program, recognizing clusters at a low case density and making informed decisions to pursue these clusters for further investigation can be difficult (Unpublished internal report, PHAC, 2014). However, the use of WGS has shown the ability to weed out false clusters in both this study and past studies (84), allowing for these decisions to be made with greater confidence. The difficulty in deciding to pursue clusters for further investigation was likely an issue with the PNC-Newp-4 cluster, and the additional cases ruled in by WGS could have made the pursuit of more information through an epidemiological investigation more appealing. The difficulty in making these decisions was also an issue during the early stages of

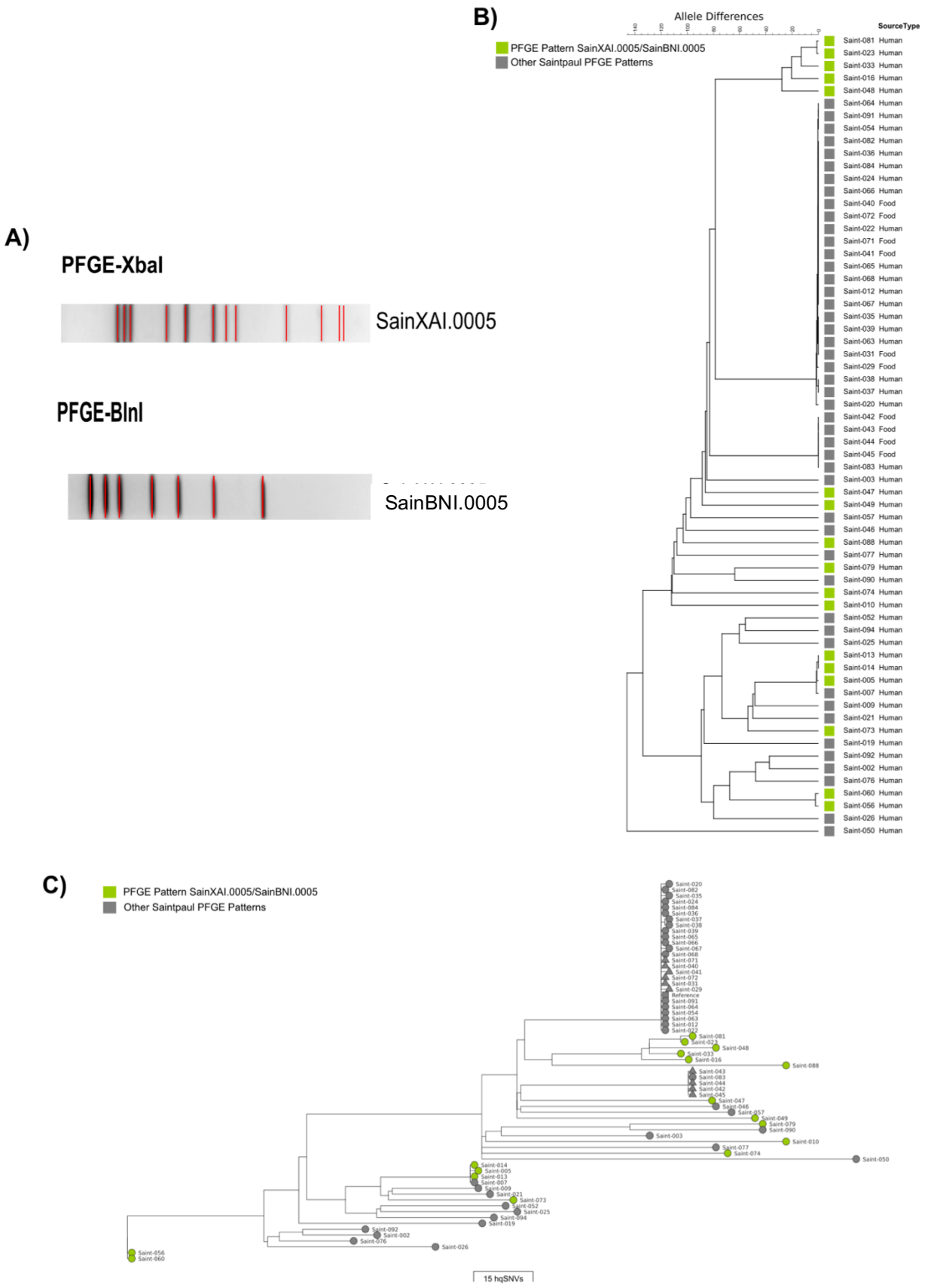
the sprouted chia outbreak investigation (Unpublished internal report, PHAC, 2014) and the improved case categorization provided by WGS-based analysis methods increases the confidence in laboratory-based results. This increased confidence improved the ability of PulseNet USA to refocus resources, and lead to an increase in solved outbreaks in comparison to when PFGE-based analysis was used (84).

4.2.5.2.1 Isolates with the Same PFGE Pattern Do Not Necessarily Form a Cluster

One important finding from the improved case categorization of WGS-based analysis is the improved discriminatory power among isolates with a shared PFGE pattern. This is especially important in assessing potential clusters of common PFGE patterns. For example, some serovars such as *Salmonella* Enteritidis show little pattern diversity, making decisions on potential clusters difficult. WGS based analysis has shown the remarkable ability to differentiate isolates from a common *Salmonella* Enteritidis PFGE pattern that is found in 40% of all Enteritidis isolates submitted to PulseNet USA. This differentiation vastly improves cluster detection and outbreak investigation among these common patterns (141).

While none of the serovars examined in this study show the same low level of pattern diversity seen among serovar Enteritidis, some common PFGE patterns were noted, such as SainXAI.0005/SainBNI.0005 (Figure 18A). This is a very common pattern combination among the clinical *Salmonella* Saintpaul isolates in Canada (data not shown) and was reported in 22% of the *Salmonella* Saintpaul isolates sequenced as part of this project. Both WGS-based analysis methods provided much higher discriminatory power in differentiating isolates with this common PFGE pattern (Figures 18B and 18C). Both methods identified three separate clusters among these isolates, indicating three separate events occurring in a single PFGE pattern. The fact that the same PFGE pattern combination can be seen in a group of isolates separated by

Figure 18: WGS-based assessment of the common PFGE pattern SainXAI.0005/SainBNI.0005 among 63 *Salmonella* Saintpaul isolates collected during the outbreak period. **(A)** DNA fingerprints of the *Xba*1 digest corresponding to pattern SainXAI.0005 and the *Bln*1 digest corresponding to pattern SainBNI.0005. **(B)** wgMLST-based tree built using character data on 4,766 loci found among these 63 isolates collected over the course of the outbreak. **(C)** SNVPhyl tree built using 1,279 hqSNVs across 93.25% of the reference genome, the SPAdes assembled genome Saint-012. Nodes in (B) and (C) are coloured to denote PFGE pattern combination.


A)
PFGE-XbaI
SainXAI.0005
PFGE-BlnI
SainBNI.0005
B)
Allele Differences
PFGE Pattern SainXAI.0005/SainBNI.0005
Other Saintpaul PFGE Patterns
SourceType
C)
PFGE Pattern SainXAI.0005/SainBNI.0005
Other Saintpaul PFGE Patterns
15 hqSNVs

many SNVs or alleles is not shocking. A PFGE-based analysis only captures a small fraction of an isolate's total genomic content, and many nucleotide changes can occur within an isolate that will have no impact on the resulting PFGE pattern (22). WGS-based data analysis platforms greatly expand on the fraction of an isolate's total genetic material which is assessed. For example, within this study, the SNVPhyl pipeline assessed genomes across a minimum of 92% of the reference genome, providing a much higher resolution on which relatedness could be determined.

4.2.5.2.2 Isolates with Different PFGE Patterns Can Still Form a Cluster

It is also important to note that isolates with different PFGE pattern combinations are able to form clusters through WGS based analysis. For example, this was seen among the *Salmonella* Saintpaul isolates from the larger of the two sprouted chia powder clusters. The 25 isolates in this larger cluster all had the same *Xba*1 pattern but displayed four different *Bln*1 patterns (Figure 19A). The differences seen in two of these *Bln*1 patterns were high enough to exclude two isolates with these *Bln*1 patterns from the investigation; however, WGS-based analysis placed these isolates into a tight cluster separated by two SNVs or alleles (Figure 19B and Figure 19C). This same phenomenon was noted by PulseNet USA during their pilot study on WGS-based analysis techniques for cluster detection of *Listeria* (84).

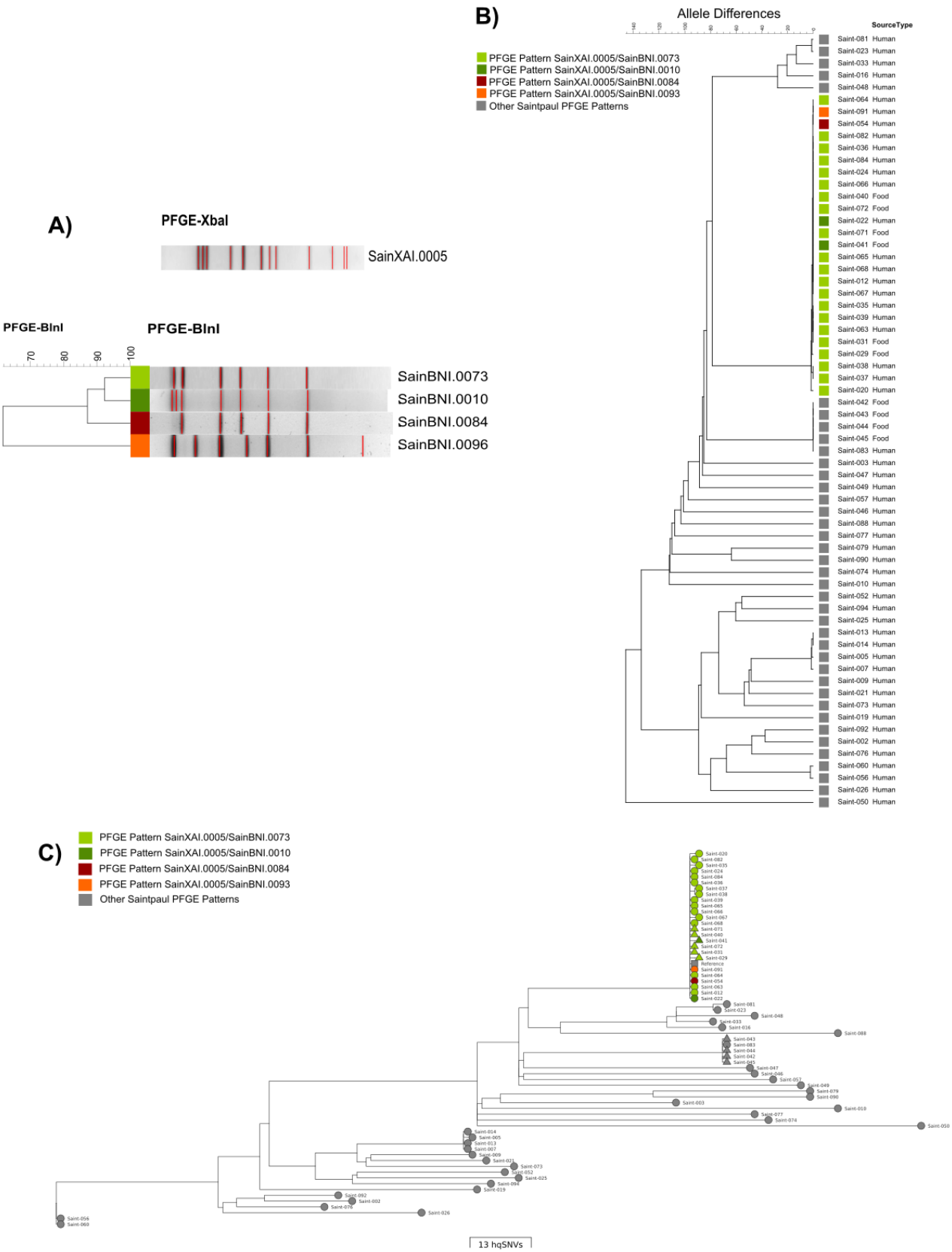
WGS-based methods compare isolates across a much larger percentage of their genetic material; however, there are still areas of the genome that are not examined through WGS-based analysis. For example, the *Salmonella* Saintpaul isolates in this study were compared across 93% of the reference genome using SNVPhyl, encompassing a total of 4.39Mbp; seven percent of the reference genome was deemed invalid and filtered by the SNVPhyl platform, and any diversity in these areas would not be reflected in the resulting phylogeny. As well, it is also important to

note that in this example, the reference genome was a draft SPAdes assembled genome, and any sequence gaps between contigs were also not sampled and therefore could not be examined. With this said, the total amount of the core genome that was not examined by SNVPhyl was quite small as the average genome size for *Salmonella* is about 4.7 Mbp in length (141). It is therefore likely that the nucleotide-level changes that resulted in the slightly different *Bln*1 patterns among these isolates were located in genomic regions which were not included in the SNVPhyl analysis, including any potential plasmids. However; the fact that these isolates with slightly different *Bln*1 patterns produced identical clustering via both SNVPhyl and wgMLST which assesses diversity over such a large area indicates these isolates are likely related and should have been included in the outbreak investigation.

4.2.5.3 WGS Does Not Negate the Importance of Additional Streams of Evidence

Although WGS data analysis in this retrospective study displayed higher resolution in comparison to PFGE-based data, its use does not negate the importance of other avenues of investigation during outbreak scenarios. Epidemiological and food safety investigations are crucial for providing the context needed for the interpretation of WGS-based data. Multiple cut-offs for determining isolate inclusion or exclusion to an outbreak event based on SNV or allele differences have been proposed in the literature (84, 130, 146, 151), and even within this retrospective investigation the cut-offs (Table 11) varied by serovar. The establishment of a single SNV or allelic numerical cut-off value to define an outbreak for all *Salmonella* serovars is not likely possible (146, 151). Much like the interpretation criteria for PFGE (61), the interpretation criteria of WGS-based data must take into account the diversity within the organism/serotype/lineage. As well, these findings should be interpreted using the context provided by all available information during an outbreak investigation.

Figure 19: Assessment of a WGS-based cluster consisting of representatives from multiple PFGE pattern combinations among 63 *Salmonella* Saintpaul isolates collected during the outbreak period. **(A)** DNA fingerprints of the *Xba*1 digest corresponding to pattern SainXAI.0005, and the *Bln*1 digests corresponding to patterns SainBNI.0073, SainBNI.0010, SainBNI.0084, and SainBNI.0096 that was among the Saintpaul isolates linked to the sprouted chia outbreak cluster. Scale indicates the percent similarity of the PFGE DNA fingerprint. **(B)** wgMLST-based tree built using character data on 4,766 loci found among these 63 isolates. **(C)** SNVPhyl tree of the isolates built using 1,279 hqSNVs across 93.25% of the reference genome, the SPAdes assembled genome Saint-012. Nodes across the figures are colour coded to denote the *Bln*1 PFGE-pattern.


A)
PFGE-XbaI
SainXAI.0005
PFGE-BlnI
PFGE-BlnI
70
80
90
100
SainBNI.0073
SainBNI.0010
SainBNI.0084
SainBNI.0096
B)
Allele Differences
SourceType
PFGE Pattern SainXAI.0005/SainBNI.0073
PFGE Pattern SainXAI.0005/SainBNI.0010
PFGE Pattern SainXAI.0005/SainBNI.0084
PFGE Pattern SainXAI.0005/SainBNI.0093
Other Saintpaul PFGE Patterns
Saint-081 Human
Saint-023 Human
Saint-033 Human
Saint-016 Human
Saint-048 Human
Saint-064 Human
Saint-091 Human
Saint-054 Human
Saint-082 Human
Saint-036 Human
Saint-084 Human
Saint-024 Human
Saint-066 Human
Saint-040 Food
Saint-072 Food
Saint-022 Human
Saint-071 Food
Saint-041 Food
Saint-065 Human
Saint-068 Human
Saint-012 Human
Saint-067 Human
Saint-035 Human
Saint-039 Human
Saint-063 Human
Saint-031 Food
Saint-029 Food
Saint-038 Human
Saint-037 Human
Saint-020 Human
Saint-042 Food
Saint-043 Food
Saint-044 Food
Saint-045 Food
Saint-083 Human
Saint-003 Human
Saint-047 Human
Saint-049 Human
Saint-057 Human
Saint-046 Human
Saint-088 Human
Saint-077 Human
Saint-079 Human
Saint-090 Human
Saint-074 Human
Saint-010 Human
Saint-052 Human
Saint-094 Human
Saint-025 Human
Saint-013 Human
Saint-014 Human
Saint-005 Human
Saint-007 Human
Saint-009 Human
Saint-021 Human
Saint-073 Human
Saint-019 Human
Saint-092 Human
Saint-002 Human
Saint-076 Human
Saint-060 Human
Saint-056 Human
Saint-026 Human
Saint-050 Human
C)
PFGE Pattern SainXAI.0005/SainBNI.0073
PFGE Pattern SainXAI.0005/SainBNI.0010
PFGE Pattern SainXAI.0005/SainBNI.0084
PFGE Pattern SainXAI.0005/SainBNI.0093
Other Saintpaul PFGE Patterns
13 hqSNVs

In the context of a multi-serovar event, epidemiological information is especially crucial when trying to connect multiple strains to a single food commodity. WGS-based analysis lacks the ability to connect multiple diverse strains that may exist in a single commodity, and other avenues of investigation are essential to provide that information. Therefore, the continued need and importance of epidemiological and food safety investigatory evidence remains crucial in the age of genomics.

4.3 Considerations for the Implementation of WGS for *Salmonella* Surveillance

The impacts of integrating genomic data into the existing *Salmonella* surveillance networks in Canada cannot be ignored. While WGS technologies present many benefits, they are also a disruptive force, and considerations must be given to the impacts this technology will have on the existing surveillance paradigm and the shifts that will occur with its adoption. Currently, the surveillance paradigm for bacterial enteric pathogens causing human disease in Canada is structured around two main programs, NESP and PulseNet Canada. These programs are separated not only by the types of tests they perform, but also the subtyping resolution these tests provide, the times at which they are performed, and the public health decisions that are made using the collected information (Figure 20A).

As Canada's national diagnostic and reference laboratories with surveillance functions continue to move forward with WGS implementation, the current surveillance paradigm will be shifted. WGS will provide the simultaneous collection of information that will feed into all surveillance programs. Surveillance program mandates will not be defined by the test type and time in which this test is completed, but instead, they will be defined by resolution of the analysis method. In a genomics-based system, low-resolution *in silico* typing information could be collected for routine identification and diagnostics from the same data set from which high-

resolution wgMLST data is generated. This paradigm shift would negate the time factor that often separates the performance of these tests in the current system (Figure 20B). However, this changing paradigm should not be used to diminish the mandates of the various surveillance programs. Information collected and stored by NESP would still be important to provide early outbreak detection, allocate investigation resources, organize PulseNet Canada databases, and inform questions raised during epidemiological investigations.

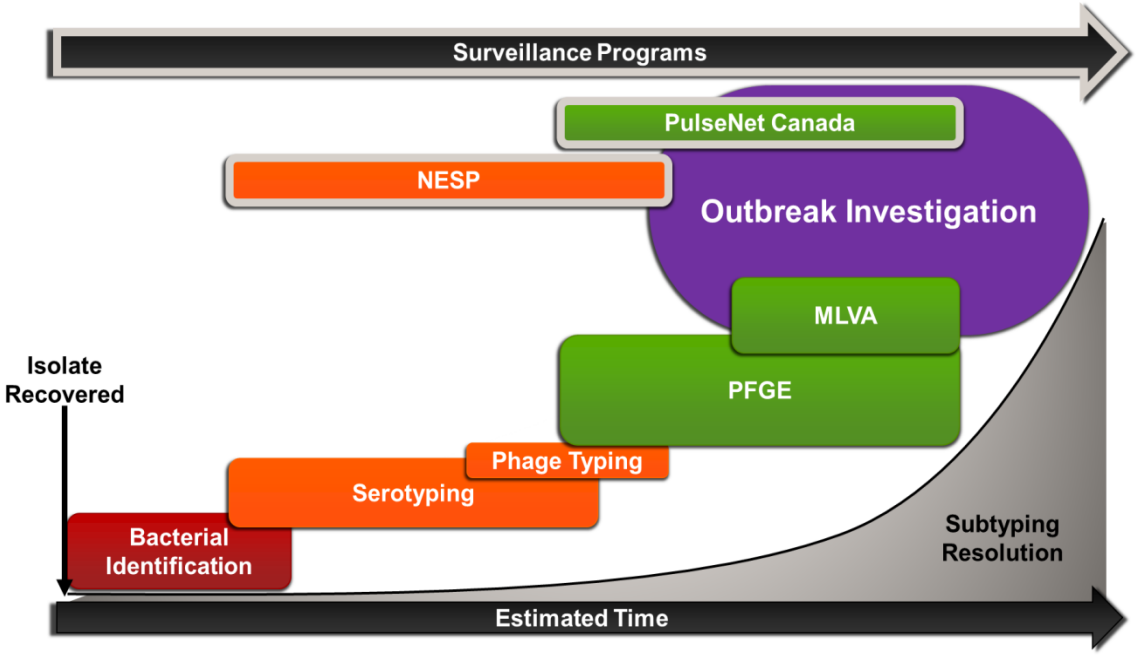
Both NESP and PulseNet Canada rely heavily on serotyping information to help carry out their mandates (27, 29). In NESP, clinical infections are recorded at the serovar level, and current disease numbers are compared to historical levels for the rapid detection of changes in disease frequencies, serving as an indicator of potential clusters (27). This information is utilized to help inform the investigation into potential clusters by PulseNet Canada, as well as organize their databases and inform their case finding activities.

This is especially true for serovars that are not routinely typed by PFGE as large increases in actual numbers over expected values would lead to an investigation (29). This is best highlighted by a 2014-2015 *Salmonella* Reading outbreak which showed a massive spike in the number of isolates reported to NESP from the yearly expected number of *Salmonella* Reading infections. Early reporting of this increase led to both a laboratory and epidemiological investigation (32). *Salmonella* serotypes can also provide important information to epidemiological investigations by providing information that may help determine potential sources (10).

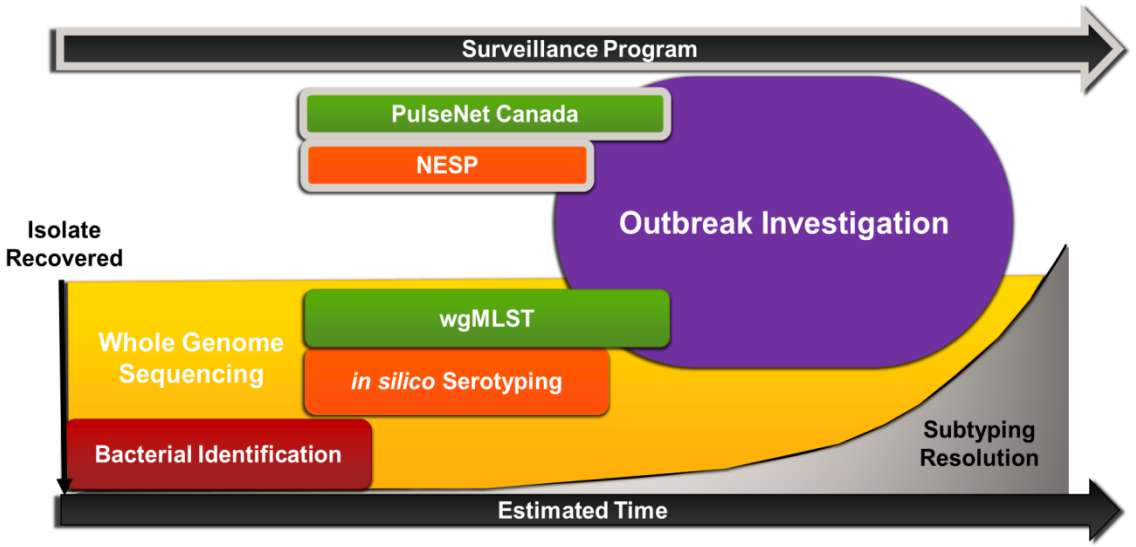
However, this changing paradigm of surveillance in Canada does point to the need for better and more seamless integration and sharing of data between NESP and PulseNet Canada as these programs will utilize the same dataset to perform their respective analyses. Down the road,

Figure 20: Paradigms of surveillance for *Salmonella enterica* and other priority bacterial pathogens in Canada. **(A)** Current surveillance paradigm encompassing multiple tests at varying levels of resolution feeding into different systems. **(B)** Proposed future surveillance paradigm encompassing a single WGS-based test capable of providing multiple levels of analysis at different levels of resolution and feeding into different systems.

A)


Surveillance Programs
PulseNet Canada
NESP
Outbreak Investigation
MLVA
PFGE
Phage Typing
Serotyping
Bacterial Identification
Isolate Recovered
Subtyping Resolution
Estimated Time

B)


Surveillance Program
PulseNet Canada
NESP
Outbreak Investigation
Isolate Recovered
Whole Genome Sequencing
wgMLST
in silico Serotyping
Bacterial Identification
Subtyping Resolution
Estimated Time

the sharing of this information with other surveillance programs, such as FoodNet Canada and the Canadian Integrated Program for Antimicrobial Resistance Surveillance would also be beneficial to ensure that various systems are not replicating the generation of expansive datasets. The Bioinformatics Core at the NML has developed a bioinformatics and data sharing platform called IRIDA that allows for data generators to maintain local ownership of sample data while facilitating data sharing with specified partners. This platform has also integrated many important bioinformatic tools for the quality control, assembly, *in silico* serotyping (via SISTR), and phylogenetic tree building (via SNVPhyl) of WGS data sets. As well, the WGS data that is stored and shared through IRIDA can be seamlessly pulled from IRIDA for analysis by other programs, including Bionumerics or the Galaxy suit of WGS tools.

Further considerations must be given to the costs associated with sequencing. Previous rough estimates of the costs incurred from WGS indicate that the total cost per isolate is high, but remains lower than the costs of serotyping and PFGE combined (Unpublished internal report, Aleisha Reimer, 2013). As well, the costs of WGS are expected to continue to decrease in the coming years (69, 73). Lastly, the informational and technical capacities of both national and partner laboratories must also be considered to ensure all stakeholders have the technical and informational capacities and understandings to fully participate in the new paradigm.

4.4 Limitations

The use of sequencing data that was generated from size selected libraries represents a major limitation of the *in silico* serotyping study. Size selected libraries displayed poor sequencing coverage over the *rfb* region, whose assembly was required for the O-antigen prediction from SISTR and SeqSero. Using the size selected data very likely reduced the number of full matches, while also inflating the number of incorrect results, for both platforms.

Unfortunately, this issue was only detected after all data had been generated; however, the limitation of this dataset did uncover an important consideration going forward, which is to ensure libraries are not size selected if the data is to be used for *in silico* serotype prediction.

One major limitation of the sprouted chia outbreak investigation was the use of the PFGE upload date for case finding. These dates represent the one consistently reported date in the PulseNet Canada databases; however, they may not always represent a true reflection of when an individual was sick. This is especially true for serovars that are not routinely subtyped via PFGE, such as the one included in the sprouted chia outbreak. In these situations, case findings could result in scenarios where two isolates that were collected months apart from each other are uploaded over the span of a few days.

4.5 Future Directions

Already, the enterics unit at the NML is starting to utilize SISTR in parallel with traditional serotyping techniques for *Salmonella*, and the complete switch-over to genomics-based serotyping is imminent. At the same time, PulseNet Canada is also starting to phase out PFGE-based subtyping for *Salmonella* and this will be replaced solely with WGS data analysis techniques. As more WGS data is generated on this organism, opportunities for in-depth research projects will develop to further help understand the biology, transmission, and virulence of this enteric pathogen.

4.6 Conclusions

WGS-based technologies will transform the current surveillance systems for *Salmonella* and other enteric pathogens, not only in Canada but around the world. However, these transformations should not be used to diminish the importance of the various surveillance programs, their mandates, and the need for multiple streams of evidence. WGS data has shown

the ability to provide accurate serotyping results for *Salmonella* using SISTR and has displayed the ability to improve the case categorization of isolates during cluster detection and outbreak investigation using appropriate clustering tools such as wgMLST and SNVPhyl.

References

1. **World Health Organization.** 2015. WHO estimates of the global burden of foodborne diseases: foodborne disease burden epidemiology reference group 2007-2015. World Health Organization. Geneva, Switzerland.
2. **Newell DG, Koopmans M, Verhoef L, Duizer E, Aidara-Kane A, Sprong H, Opsteegh M, Langelaar M, Threfall J, Scheutz F, van der Giessen J, Kruse H.** 2010. Food-borne diseases - the challenges of 20 years ago still persist while new ones continue to emerge. Int J Food Microbiol **139 (Supp 1):**S3-15.
3. **Thomas MK, Murray R, Flockhart L, Pintar K, Pollari F, Fazil A, Nesbitt A, Marshall B.** 2013. Estimates of the burden of foodborne illness in Canada for 30 specified pathogens and unspecified agents, circa 2006. Foodborne Pathog Dis **10:**639-648.
4. **Haagsma JA, Polinder S, Stein CE, Havelaar AH.** 2013. Systematic review of foodborne burden of disease studies: quality assessment of data and methodology. Int J Food Microbiol **166:**34-47.
5. **Thomas MK, Murray R, Flockhart L, Pintar K, Fazil A, Nesbitt A, Marshall B, Tataryn J, Pollari F.** 2015. Estimates of foodborne illness-related hospitalizations and deaths in Canada for 30 specified pathogens and unspecified agents. Foodborne Pathog Dis **12:**820-827.
6. **Majowicz SE, McNab WB, Sockett P, Henson TS, Dore K, Edge VL, Buffett MC, Fazil A, Read S, McEwen S, Stacey D, Wilson JB.** 2006. Burden and cost of gastroenteritis in a Canadian community. J Food Prot **69:**651-659.
7. **Henson SJ, Majowicz SE, Masakure O, Sockett PN, MacDougall L, Edge VL, Thomas MK, Fyfe M, Kovacs SJ, Jones AQ.** 2008. Estimation of the costs of acute gastrointestinal illness in British Columbia, Canada. Int J Food Microbiol **127:**43-52.
8. **Thomas MK, Vriezen R, Farber JM, Currie A, Schlech W, Fazil A.** 2015. Economic Cost of a *Listeria monocytogenes* Outbreak in Canada, 2008. Foodborne Pathog Dis **12:**966-971.
9. **Nataro JP, Bopp CA, Fields PI, Kaper JB, Strockbine NA.** 2011. *Escherichia*, *Shigella*, and *Salmonella*, p 603-626. *In* Versalovic J, Carroll KC, Jorgensen JH, Funke G, Landry ML, DW W (ed), Manual of clinical microbiology, 10th ed, vol 1. ASM Press, Washington, D.C.
10. **Jackson BR, Griffin PM, Cole D, Walsh KA, Chai SJ.** 2013. Outbreak-associated *Salmonella enterica* serotypes and food Commodities, United States, 1998-2008. Emerg Infect Dis **19:**1239-1244.
11. **Tindall BJ, Grimont PA, Garrity GM, Euzeby JP.** 2005. Nomenclature and taxonomy of the genus *Salmonella*. Int J Syst Evol Microbiol **55:**521-524.
12. **Le Minor L, Popoff MY.** 1987. Designation of *Salmonella enterica* sp. nov., nom. rev., as the Type and Only Species of the Genus *Salmonella*. Int J Syst Bacter **37:**465-468.
13. **Grimont PAD, Weill F-X.** 2007. Antigenic formulae of the *Salmonella* serovars. WHO Collaborating Center for Reference and Research on *Salmonella*. Paris, France.
14. **Gal-Mor O, Boyle EC, Grassl GA.** 2014. Same species, different diseases: how and why typhoidal and non-typhoidal *Salmonella enterica* serovars differ. Front Microbiol **5:**391.
15. **Li H, Wang H, apos, Aoust J-Y, Maurer J.** 2013. *Salmonella* Species†, p 225-261. *In*

Doyle MP, Buchanan RL (ed), Food microbiology: Fundamentals and frontiers, 4th ed. ASM Press, Washington, D.C.
16. **Ricke SC, Ok-Kyung K, Foley S, Nayak R.** 2013. *Salmonella*, p 112-137. *In* Labbe RG, Garcia S (ed), Guide to foodborne pathogens, 2nd ed. John Wiley & Sons, Hoboken, NJ.
17. **Podolak R, Enache E, Stone W, Black DG, Elliott PH.** 2010. Sources and risk factors for contamination, survival, persistence, and heat resistance of *Salmonella* in low-moisture foods. J Food Prot **73:**1919-1936.
18. **Haraga A, Ohlson MB, Miller SI.** 2008. Salmonellae interplay with host cells. Nat Rev Microbiol **6:**53-66.
19. **Dey M, Mayo JA, Saville D, Wolyniak C, Klontz KC.** 2013. Recalls of foods due to microbiological contamination classified by the U.S. Food and Drug Administration, fiscal years 2003 through 2011. J Food Prot **76:**932-938.
20. **Centers for Disease Control and Prevention.** 2009. Multistate outbreak of *Salmonella* infections associated with peanut butter and peanut butter-containing products - United States, 2008-2009. Morb Mortal Wkly Rep **58:**85-90.
21. **Mank L, Mandour M, Rabatsky-Ehr T, Phan Q, Krasnitski J, Brockmeyer J, Bushnell L, Applewhite C, Cartter M, Kattan J.** 2010. Multiple-serotype *Salmonella* gastroenteritis outbreak after a reception - Connecticut, 2009. MMWR Morb Mortal Wkly Rep **59:**1093-1097.
22. **Barrett TJ, Gerner-Smidt P, Swaminathan B.** 2006. Interpretation of pulsed-field gel electrophoresis patterns in foodborne disease investigations and surveillance. Foodborne Pathog Dis **3:**20-31.
23. **Paine S, Thornley C, Wilson M, Dufour M, Sexton K, Miller J, King G, Bell S, Bandaranayake D, Mackereth G.** 2014. An outbreak of multiple serotypes of *Salmonella* in New Zealand linked to consumption of contaminated tahini imported from Turkey. Foodborne Pathog Dis **11:**887-892.
24. **Parmley EJ, Pintar K, Majowicz S, Avery B, Cook A, Jokinen C, Gannon V, Lapen DR, Topp E, Edge TA, Gilmour M, Pollari F, Reid-Smith R, Irwin R.** 2013. A Canadian application of one health: integration of *Salmonella* data from various Canadian surveillance programs (2005-2010). Foodborne Pathog Dis **10:**747-756.
25. **Government of Canada.** 2014. 2013 Short Report - FoodNet Canada: Canada's National Integrated Enteric Pathogen Surveillance System. Public Health Agency of Canada Ottawa, ON. http://publications.gc.ca/pub?id=470245&sl=0.
26. **Vik J, Hexemer A.** 2014. Summary: Canada's food-borne illness outbreak response protocol. Can Communicable Dis Rep **40:**306-310.
27. **Government of Canada.** 2014. National Enteric Surveillance Program (NESP): Annual Summary 2012. Public Health Agency of Canada, Guelph, ON. http://publications.gc.ca/pub?id=465160&sl=0.
28. **Government of Canada.** 2016. National Enteric Surveillance Program (NESP): Annual Summary 2014. Public Health Agency of Canada, Guelph, ON. http://publications.gc.ca/collections/collection_2016/aspc-phac/HP37-15-2014-eng.pdf.
29. **Swaminathan B, Gerner-Smidt P, Ng LK, Lukinmaa S, Kam KM, Rolando S, Gutierrez 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.

30. **Rumore JL, Tschetter L, Nadon C.** 2016. The Impact of Multilocus Variable-Number Tandem-Repeat Analysis on PulseNet Canada *Escherichia coli* O157:H7 Laboratory Surveillance and Outbreak Support, 2008-2012. Foodborne Pathog Dis **13:**255-261.
31. **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 Suppl 2:**57-62.
32. **Tanguay F, Vrbova L, Anderson M, Whitfield Y, Macdonald L, Tschetter L, Hexemer A, Salmonella Reading Investigation Team.** 2017. Outbreak of *Salmonella* Reading in persons of Eastern Mediterranean origin in Canada, 2014-2015. Can Commun Dis Rep **42:**14-20.
33. **Landry L, Phan Q, Kelly S, Phillips K, Onofrey S, Daly ER, Deasy M, Spayne M, Lynch M, Olson CK.** 2007. *Salmonella* Oranienburg infections associated with fruit salad served in health-care facilities--northeastern United States and Canada, 2006. MMWR Morb Mortal Wkly Rep **56:**1025-1028.
34. **Deeks S, Ellis A, Ciebin B, Khakhria R, Naus M, Hockin J.** 1998. *Salmonella* Oranienburg, Ontario. Can Commun Dis Rep **24:**177-179.
35. **Mody RK, Greene SA, Gaul L, Sever A, Pichette S, Zambrana I, Dang T, Gass A, Wood R, Herman K, Cantwell LB, Falkenhorst G, Wannemuehler K, Hoekstra RM, McCullum I, Cone A, Franklin L, Austin J, Delea K, Behravesh CB, Sodha SV, Yee JC, Emanuel B, Al-Khaldi SF, Jefferson V, Williams IT, Griffin PM, Swerdlow DL.** 2011. National outbreak of *Salmonella* serotype Saintpaul infections: importance of Texas restaurant investigations in implicating jalapeno peppers. PLoS One **6:**e16579.
36. **Basler C, Forshey TM, Machesky K, Erdman MC, Gomez TM, Nguyen TA, Behravesh CB.** 2014. Multistate outbreak of human *Salmonella* infections linked to live poultry from a mail-order hatchery in Ohio--March-September 2013. MMWR Morb Mortal Wkly Rep **63:**222.
37. **Scharff RL, Besser J, Sharp DJ, Jones TF, Peter GS, Hedberg CW.** 2016. An Economic Evaluation of PulseNet: A Network for Foodborne Disease Surveillance. Am J Prev Med **50:**S66-73.
38. **Gerner-Smidt P, Hyytiä-trees E, Rota PA.** 2011. Molecular Epidemiology, p 100-123. *In* Versalovic J, Carroll KC, Jorgensen JH, Funke G, Landry ML, DW W (ed), Manual of clinical microbiology, 10th ed, vol 1. ASM Press, Washington, D.C.
39. **Kim S, Frye JG, Hu J, Fedorka-Cray PJ, Gautom R, Boyle DS.** 2006. Multiplex PCR-based method for identification of common clinical serotypes of *Salmonella enterica* subspecies *enterica* J Clin Microbiol **44:**3608-3615.
40. **Wattiau P, Boland C, Bertrand S.** 2011. Methodologies for *Salmonella enterica* subsp. *enterica* subtyping: gold standards and alternatives. Appl Environ Microbiol **77:**7877-7885.
41. **Yoshida C, Kruczkiewicz P, Laing CR, Lingohr EJ, Gannon VPJ, Nash JHE, Taboada EN.** 2016. The *Salmonella In Silico* Typing Resource (SISTR): an open web-accessible tool for rapidly typing and subtyping draft *Salmonella* genome assemblies. PLoS One **11:**e0147101.
42. **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:**1685-1692.
43. **Salazar JK, Wang Y, Yu S, Wang H, Zhang W.** 2015. Polymerase chain reaction-

based serotyping of pathogenic bacteria in food. J Microbiol Methods **110:**18-26.
44. **Reyes RE, Gonzalez CR, Jimenez RC, Herrera MO, Andrade AA.** 2012. Mechanism of O-antigen structural variation of bacterial lipopolysaccharide (LPS). *In* Karunaratne DN (ed), The Complex World of Polysaccharides doi:10.5772/48147. InTech, Online.
45. **Switt AM, Sulakvelidze A, Wiedmann M, Kropinski A, Wishart D, Poppe C, Liang Y.** 2015. *Salmonella* Phages and Prophages: Genomics, Taxonomy, and Applied Aspects, p 237-287. *In* Schatten H, Eisenstark A (ed), Salmonella: Methods and Protocols. Springer New York, New York, NY.
46. **Yamamoto S, Kutsukake K.** 2006. FljA-mediated posttranscriptional control of phase 1 flagellin expression in flagellar phase variation of *Salmonella enterica* serovar Typhimurium. J Bacteriol **188:**958-967.
47. **Hendriksen RS, Vieira AR, Karlsmose S, Lo Fo Wong DM, Jensen AB, Wegener HC, Aarestrup FM.** 2011. Global monitoring of *Salmonella* serovar distribution from the World Health Organization Global Foodborne Infections Network Country Data Bank: results of quality assured laboratories from 2001 to 2007. Foodborne Pathog Dis **8:**887-900.
48. **Hendriksen RS, Mikoleit M, Carlson VP, Karlsmose S, Vieira AR, Jensen AB, Seyfarth AM, DeLong SM, Weill FX, Lo Fo Wong DM, Angulo FJ, Wegener HC, Aarestrup FM.** 2009. WHO Global Salm-Surv external quality assurance system for serotyping of *Salmonella* isolates from 2000 to 2007. J Clin Microbiol **47:**2729-2736.
49. **Wattiau P, Van Hessche M, Schlicker C, Vander Veken H, Imberechts H.** 2008. Comparison of classical serotyping and PremiTest assay for routine identification of common *Salmonella enterica* serovars. J Clin Microbiol **46:**4037-4040.
50. **Yoshida C, Lingohr EJ, Trognitz F, MacLaren N, Rosano A, Murphy SA, Villegas A, Polt M, Franklin K, Kostic T, Kropinski AM, Card RM.** 2014. Multi-laboratory evaluation of the rapid genoserotyping array (SGSA) for the identification of *Salmonella* serovars. Diagn Microbiol Infect Dis **80:**185-190.
51. **Herrera-Leon S, Ramiro R, Arroyo M, Diez R, Usera MA, Echeita MA.** 2007. Blind comparison of traditional serotyping with three multiplex PCRs for the identification of *Salmonella* serotypes. Res Microbiol **158:**122-127.
52. **Nesbitt A, Ravel A, Murray R, McCormick R, Savelli C, Finley R, Parmley J, Agunos A, Majowicz SE, Gilmour M.** 2012. Integrated surveillance and potential sources of *Salmonella* Enteritidis in human cases in Canada from 2003 to 2009. Epidemiol Infect **140:**1757-1772.
53. **Demczuk W, Soule G, Clark C, Ackermann HW, Easy R, Khakhria R, Rodgers F, Ahmed R.** 2003. Phage-based typing scheme for *Salmonella enterica* serovar Heidelberg, a causative agent of food poisonings in Canada. J Clin Microbiol **41:**4279-4284.
54. **Zheng J, Pettengill J, Strain E, Allard MW, Ahmed R, Zhao S, Brown EW.** 2014. Genetic diversity and evolution of *Salmonella enterica* serovar Enteritidis strains with different phage types. J Clin Microbiol **52:**1490-1500.
55. **Baggesen DL, Sorensen G, Nielsen EM, Wegener HC.** 2010. Phage typing of *Salmonella* Typhimurium - is it still a useful tool for surveillance and outbreak investigation? Euro Surveill **15:**19471.
56. **Gerner-Smidt P, Hise K, Kincaid J, Hunter S, Rolando S, Hyytia-Trees E, Ribot EM, Swaminathan B.** 2006. PulseNet USA: a five-year update. Foodborne Pathog Dis

3:9-19.
57. **Goering RV.** 2010. Pulsed field gel electrophoresis: a review of application and interpretation in the molecular epidemiology of infectious disease. Infect Genet Evol **10:**866-875.
58. **Hedberg CW, Besser JM.** 2006. Commentary: cluster evaluation, PulseNet, and public health practice. Foodborne Pathog Dis **3:**32-35.
59. **Camarda A, Circella E, Pupillo A, Legretto M, Marino M, Pugliese N.** 2015. Pulsed-field gel electrophoresis of *Salmonella enterica*. Methods Mol Biol **1301:**191-210.
60. **Tenover FC, Arbeit RD, Goering RV, Mickelsen PA, Murray BE, Persing DH, Swaminathan B.** 1995. Interpreting chromosomal DNA restriction patterns produced by pulsed-field gel electrophoresis: criteria for bacterial strain typing. J Clin Microbiol **33:**2233-2239.
61. **Government of Canada.** 2011. Weight of evidence: factors to consider for appropriate and timely action in a foodborne illness outbreak investigation. Health Canada, Public Health Agency of Canada, Canadian Food Inspection Agency, Ottawa, ON. . http://www.hc-sc.gc.ca/fn-an/alt_formats/pdf/pubs/securit/2011-food-illness-outbreak-eclosion-malad-ailments-eng.pdf.
62. **Nadon CA, Trees E, Ng LK, Moller Nielsen E, Reimer A, Maxwell N, Kubota KA, Gerner-Smidt P.** 2013. Development and application of MLVA methods as a tool for inter-laboratory surveillance. Euro Surveill **18:**20565.
63. **Maiden MC, Bygraves JA, Feil E, Morelli G, Russell JE, Urwin R, Zhang Q, Zhou J, Zurth K, Caugant DA, Feavers IM, Achtman M, Spratt BG.** 1998. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. Proc Natl Acad Sci USA **95:**3140-3145.
64. **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:**1355-1361.
65. **Jolley KA, Maiden MC.** 2014. Using MLST to study bacterial variation: prospects in the genomic era. Future Microbiol **9:**623-630.
66. **Kidgell C, Reichard U, Wain J, Linz B, Torpdahl M, Dougan G, Achtman M.** 2002. *Salmonella* Typhi, the causative agent of typhoid fever, is approximately 50,000 years old. Infect Genet Evol **2:**39-45.
67. **Sangal V, Harbottle H, Mazzoni CJ, Helmuth R, Guerra B, Didelot X, Paglietti B, Rabsch W, Brisse S, Weill FX, Roumagnac P, Achtman M.** 2010. Evolution and population structure of *Salmonella enterica* serovar Newport. J Bacteriol **192:**6465-6476.
68. **Gilmour MW, Graham M, Reimer A, Van Domselaar G.** 2013. Public health genomics and the new molecular epidemiology of bacterial pathogens. Public Health Genomics **16:**25-30.
69. **Ronholm J, Nasheri N, Petronella N, Pagotto F.** 2016. Navigating microbiological food safety in the era of whole-genome sequencing. Clin Microbiol Rev **29:**837-857.
70. **Metzker ML.** 2010. Sequencing technologies - the next generation. Nat Rev Genet **11:**31-46.
71. **Douglas GL, Pfeiler E, Duong T, Klaenhammer TR.** 2013. Genomics and Proteomics of Foodborne Microorganisms, p 975-996. *In* Doyle MP, Buchanan RL (ed), Food microbiology: Fundamentals and frontiers, 4th ed. ASM Press, Washington, D.C.
72. **Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, Bertoni A,**

Swerdlow HP, Gu Y. 2012. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics **13:**341.

73. **Lynch T, Petkau A, Knox N, Graham M, Van Domselaar G.** 2016. A primer of infectious disease bacterial genomics. Clin Microbiol Rev **29:**881-913.
74. **Petkau A.** 2015. IRIDA: A genomic epidemiology platform built on top of Galaxy, Galaxy Community Conference 2015 United Kingdom July 8 2015. http://www.irida.ca/wp-content/uploads/2014/06/IRIDAGCC2015Presentation.pdf
75. **Wood DE, Salzberg SL.** 2014. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol **15:**R46.
76. **Achtman M, Wain J, Weill FX, Nair S, Zhou Z, Sangal V, Krauland MG, Hale JL, Harbottle H, Uesbeck A, Dougan G, Harrison LH, Brisse S.** 2012. Multilocus sequence typing as a replacement for serotyping in *Salmonella enterica*. PLoS Pathog **8:**e1002776.
77. **Ashton PM, Nair S, Peters T, Bale J, Powell DG, Painset A, Tewolde R, Schaefer U, Jenkins C, T.J. D, dePinna EM, Grant KA, Salmonella Whole Genome Sequencing Implementation Group.** 2016. Identification and typing of *Salmonella* for public health surveillance using whole genome sequencing. PeerJ **4:**e1752.
78. **Issenhuth-Jeanjean S, Roggentin P, Mikoleit M, Guibourdenche M, de Pinna E, Nair S, Fields PI, Weill FX.** 2014. Supplement 2008-2010 (no. 48) to the White-Kauffmann-Le Minor scheme. Res Microbiol **165:**526-530.
79. **Inouye M, Dashnow H, Raven LA, Schultz MB, Pope BJ, Tomita T, Zobel J, Holt KE.** 2014. SRST2: Rapid genomic surveillance for public health and hospital microbiology labs. Genome Med **6:**90.
80. **Maiden MC, Jansen van Rensburg MJ, Bray JE, Earle SG, Ford SA, Jolley KA, McCarthy ND.** 2013. MLST revisited: the gene-by-gene approach to bacterial genomics. Nat Rev Microbiol **11:**728-736.
81. **McQuiston JR, Parrenas R, Ortiz-Rivera M, Gheesling L, Brenner F, Fields PI.** 2004. Sequencing and comparative analysis of flagellin genes *fliC*, *fljB*, and *flpA* from *Salmonella*. J Clin Microbiol **42:**1923-1932.
82. **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:**6099-6105.
83. **Reeves PR, Cunneen MM, Liu B, Wang L.** 2013. Genetics and evolution of the *Salmonella* galactose-initiated set of O antigens. PLoS One **8:**e69306.
84. **Jackson BR, Tarr C, Strain E, Jackson KA, Conrad A, Carleton H, Katz LS, Stroika S, Gould LH, Mody RK, Silk BJ, Beal J, Chen Y, Timme R, Doyle M, Fields A, Wise M, Tillman G, Defibaugh-Chavez S, Kucerova Z, Sabol A, Roache K, Trees E, Simmons M, Wasilenko J, Kubota K, Pouseele H, Klimke W, Besser J, Brown E, Allard M, Gerner-Smidt P.** 2016. Implementation of nationwide real-time whole-genome sequencing to enhance listeriosis outbreak detection and investigation. Clin Infect Dis **63:**380-386.
85. **Nadon C, Van Walle I, Gerner-Smidt P, Campos J, Chinen I, Concepcio-Acevedo J, Gilpin B, Smith AM, Kam KM, Perez E, Kubota K, Takkinen J, Nielsen EM, Carleton H, FWD-NEXT Expert Panel.** 2017. Puslenet International: Vision for the Implementation of Whole Genome Sequencing (WGS) for Global Food-borne Disease

Surveillance. Euro Surveill **22:**30544.
86. **Petkau A, Mabon P, Sieffert C, Knox N, Cabral J, Iskander M, Iskander M, Weedmark K, Zaheer R, Katz LS, Nadon C, Reimer A, Taboada E, Beiko RG, Hsiao W, Brinkman F, Graham M, Van Domselaar G.** 2017. SNVPhyl: A Single Nucleotide Variant Phylogenomics pipeline for microbial genomic epidemiology. Microbial Genomics **3:** doi:10.1099/mgen.1090.000116.
87. **Guindon S, Gascuel O.** 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol **52:**696-704.
88. **Baldauf SL.** 2003. Phylogeny for the faint of heart: a tutorial. Trends Genet **19:**345-351.
89. **Chen Y, Luo Y, Carleton H, Timme R, Melka D, Muruvanda T, Wang C, Kastanis G, Katz LS, Turner L, Fritzinger A, Moore T, Stones R, Blankenship J, Salter M, Parish M, Hammack TS, Evans PS, Tarr CL, Allard MW, Strain EA, Brown EW.** 2017. Whole genome and core genome multilocus sequence typing and single nucleotide polymorphism analyses of *Listeria monocytogenes* associated with an outbreak linked to cheese, United States, 2013. Appl Environ Microbiol doi:10.1128/aem.00633-17.
90. **BioNumerics.** 2016. *Salmonella enterica* schema for whole genome sequence typing. Applied Maths. Beligium. http://www.applied-maths.com/sites/default/files/extra/Release-Note-Salmonella-enterica-schema.pdf
91. **Yoshida C, Gurnik S, Ahmad A, Blimkie T, Murphy SA, Kropinski AM, Nash JH.** 2016. Evaluation of molecular methods for the identification of *Salmonella* serovars. J Clin Microbiol **54:**1992-1998.
92. **Magoc T, Salzberg SL.** 2011. FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics **27:**2957-2963.
93. **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:**455-477.
94. **Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL.** 2009. BLAST+: architecture and applications. BMC Bioinformatics **10:**421.
95. **Cock PJ, Gruning BA, Paszkiewicz K, Pritchard L.** 2013. Galaxy tools and workflows for sequence analysis with applications in molecular plant pathology. PeerJ **1:**e167.
96. **Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R.** 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics **25:**2078-2079.
97. **Milne I, Stephen G, Bayer M, Cock PJA, Pritchard L, Cardle L, Shaw PD, Marshall D.** 2013. Using Tablet for visual exploration of second-generation sequencing data. Briefings in Bioinformatics **14:**193-202.
98. **Quinlan AR, Hall IM.** 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics **26:**841-842.
99. **Mackinnon A.** 2000. A spreadsheet for the calculation of comprehensive statistics for the assessment of diagnostic tests and inter-rater agreement. Comput Biol Med **30:**127-134.
100. **Boland C, Bertrand S, Mattheus W, Dierick K, Jasson V, Rosseel T, Van Borm S, Mahillon J, Wattiau P.** 2015. Extensive genetic variability linked to IS26 insertions in the *fljB* promoter region of atypical monophasic variants of *Salmonella enterica* serovar Typhimurium. Appl Environ Microbiol **81:**3169-3175.
101. **Toboldt A, Tietze E, Helmuth R, Junker E, Fruth A, Malorny B.** 2013. Population

structure of *Salmonella enterica* serovar 4,[5],12:b:- strains and likely sources of human infection. Appl Environ Microbiol **79:**5121-5129.
102. **Malorny B, Bunge C, Helmuth R.** 2003. Discrimination of d-tartrate-fermenting and -nonfermenting *Salmonella enterica* subsp. *enterica* isolates by genotypic and phenotypic methods. J Clin Microbiol **41:**4292-4297.
103. **Hernandez J, Bonnedahl J, Waldenstrom J, Palmgren H, Olsen B.** 2003. Salmonella in birds migrating through Sweden. Emerg Infect Dis **9:**753-755.
104. **Kaibu H, Iida K, Ueki S, Ehara H, Simasaki Y, Anzai H, Toku Y, Shirono S.** 2006. Salmonellosis of infants presumably originating from an infected turtle in Nagasaki, Japan. Jpn J Infect Dis **59:**281.
105. **Kim S, Kim SH, Kim J, Shin JH, Lee BK, Park MS.** 2011. Occurrence and distribution of various genetic structures of class 1 and class 2 integrons in *Salmonella enterica* isolates from foodborne disease patients in Korea for 16 years. Foodborne Pathog Dis **8:**319-324.
106. **Government of Canada.** August 13, 2014. Public Health Notice - Outbreak of Salmonella infections related to sprouted chia seed powder. https://www.canada.ca/en/public-health/services/public-health-notices/2014/public-health-notice-outbreak-salmonella-infections-related-sprouted-chia-seed-powder.html. Accessed June 2017.
107. **Harvey RR, Heiman Marshall KE, Burnworth L, Hamel M, Tataryn J, Cutler J, Meghnath K, Wellman A, Irvin K, Isaac L, Chau K, Locas A, Kohl J, Huth PA, Nicholas D, Traphagen E, Soto K, Mank L, Holmes-Talbot K, Needham M, Barnes A, Adcock B, Honish L, Chui L, Taylor M, Gaulin C, Bekal S, Warshawsky B, Hobbs L, Tschetter LR, Surin A, Lance S, Wise ME, Williams I, Gieraltowski L.** 2017. International outbreak of multiple *Salmonella* serotype infections linked to sprouted chia seed powder - USA and Canada, 2013-2014. Epidemiol Infect **145:**1535-1544.
108. **Nair S, Ashton P, Doumith M, Connell S, Painset A, Mwaigwisya S, Langridge G, de Pinna E, Godbole G, Day M.** 2016. WGS for surveillance of antimicrobial resistance: a pilot study to detect the prevalence and mechanism of resistance to azithromycin in a UK population of non-typhoidal *Salmonella*. J Antimicrob Chemother **71:**3400-3408.
109. **Lan JH, Yin Y, Reed EF, Moua K, Thomas K, Zhang Q.** 2015. Impact of three Illumina library construction methods on GC bias and HLA genotype calling. Hum Immunol **76:**166-175.
110. **Tyler AD, Christianson S, Knox NC, Mabon P, Wolfe J, Van Domselaar G, Graham MR, Sharma MK.** 2016. Comparison of Sample Preparation Methods Used for the Next-Generation Sequencing of *Mycobacterium tuberculosis*. PLoS One **11:**e0148676.
111. **Aigrain L, Gu Y, Quail MA.** 2016. Quantitation of next generation sequencing library preparation protocol efficiencies using droplet digital PCR assays - a systematic comparison of DNA library preparation kits for Illumina sequencing. BMC Genomics **17:**458.
112. **Head SR, Komori HK, LaMere SA, Whisenant T, Van Nieuwerburgh F, Salomon DR, Ordoukhanian P.** 2014. Library construction for next-generation sequencing: overviews and challenges. Biotechniques **56:**61-77.
113. **Franklin K, Lingohr EJ, Yoshida C, Anjum M, Bodrossy L, Clark CG, Kropinski AM, Karmali MA.** 2011. Rapid genoserotyping tool for classification of *Salmonella*

serovars. J Clin Microbiol **49:**2954-2965.
114. **Kong Q, Yang J, Liu Q, Alamuri P, Roland KL, Curtiss R.** 2011. Effect of Deletion of Genes Involved in Lipopolysaccharide Core and O-Antigen Synthesis on Virulence and Immunogenicity of *Salmonella enterica* Serovar Typhimurium. Infect Immun **79:**4227-4239.
115. **Soyer Y, Moreno Switt A, Davis MA, Maurer J, McDonough PL, Schoonmaker-Bopp DJ, Dumas NB, Root T, Warnick LD, Grohn YT, Wiedmann M.** 2009. *Salmonella enterica* serotype 4,5,12:i:-, an emerging *Salmonella* serotype that represents multiple distinct clones. J Clin Microbiol **47:**3546-3556.
116. **Switt AI, Soyer Y, Warnick LD, Wiedmann M.** 2009. Emergence, distribution, and molecular and phenotypic characteristics of *Salmonella enterica* serotype 4,5,12:i. Foodborne Pathog Dis **6:**407-415.
117. **Petrovska L, Mather AE, AbuOun M, Branchu P, Harris SR, Connor T, Hopkins KL, Underwood A, Lettini AA, Page A, Bagnall M, Wain J, Parkhill J, Dougan G, Davies R, Kingsley RA.** 2016. Microevolution of Monophasic *Salmonella* Typhimurium during Epidemic, United Kingdom, 2005-2010. Emerg Infect Dis **22:**617-624.
118. **Fookes M, Schroeder GN, Langridge GC, Blondel CJ, Mammina C, Connor TR, Seth-Smith H, Vernikos GS, Robinson KS, Sanders M, Petty NK, Kingsley RA, Baumler AJ, Nuccio SP, Contreras I, Santiviago CA, Maskell D, Barrow P, Humphrey T, Nastasi A, Roberts M, Frankel G, Parkhill J, Dougan G, Thomson NR.** 2011. *Salmonella bongori* provides insights into the evolution of the Salmonellae. PLoS Pathog **7:**e1002191.
119. **Desai PT, Porwollik S, Long F, Cheng P, Wollam A, Clifton SW, Weinstock GM, McClelland M.** 2013. Evolutionary Genomics of *Salmonella enterica* Subspecies. mBio **4:**e00579-00512.
120. **Timme RE, Pettengill JB, Allard MW, Strain E, Barrangou R, Wehnes C, Van Kessel JS, Karns JS, Musser SM, Brown EW.** 2013. Phylogenetic diversity of the enteric pathogen *Salmonella enterica* subsp. *enterica* inferred from genome-wide reference-free SNP characters. Genome Biol Evol **5:**2109-2123.
121. **McQuiston JR, Fields PI, Tauxe RV, Logsdon JM, Jr.** 2008. Do *Salmonella* carry spare tyres? Trends Microbiol **16:**142-148.
122. **Prager R, Rabsch W, Streckel W, Voigt W, Tietze E, Tschape H.** 2003. Molecular properties of *Salmonella enterica* serotype Paratyphi B distinguish between its systemic and its enteric pathovars. J Clin Microbiol **41:**4270-4278.
123. **Connor TR, Owen SV, Langridge G, Connell S, Nair S, Reuter S, Dallman TJ, Corander J, Tabing KC, Le Hello S, Fookes M, Doublet B, Zhou Z, Feltwell T, Ellington MJ, Herrera S, Gilmour M, Cloeckaert A, Achtman M, Parkhill J, Wain J, De Pinna E, Weill FX, Peters T, Thomson N.** 2016. What's in a Name? Species-Wide Whole-Genome Sequencing Resolves Invasive and Noninvasive Lineages of *Salmonella enterica* Serotype Paratyphi B. MBio **7:**e00527-00516.
124. **Litrup E, Torpdahl M, Malorny B, Huehn S, Christensen H, Nielsen EM.** 2010. Association between phylogeny, virulence potential and serovars of *Salmonella enterica*. Infect Genet Evol **10:**1132-1139.
125. **Lay KS, Vuthy Y, Song P, Phol K, Sarthou JL.** 2011. Prevalence, numbers and antimicrobial susceptibilities of *Salmonella* serovars and *Campylobacter* spp. in retail poultry in Phnom Penh, Cambodia. J Vet Med Sci **73:**325-329.

126. **Ta YT, Nguyen TT, To PB, Pham da X, Le HT, Thi GN, Alali WQ, Walls I, Doyle MP.** 2014. Quantification, serovars, and antibiotic resistance of *Salmonella* isolated from retail raw chicken meat in Vietnam. J Food Prot **77:**57-66.
127. **Wieczorek K, Osek J.** 2013. Prevalence and characterization of *Salmonella* in slaughtered cattle and beef in Poland. Bull Vet Inst Pulawy **57:**607-611.
128. **Mikoleit M, Van Duyne MS, Halpin J, McGlinchey B, Fields PI.** 2012. Variable expression of O:61 in *Salmonella* group C2. J Clin Microbiol **50:**4098-4099.
129. **Cook KA, Dobbs TE, Hlady WG, Wells JG, Barrett TJ, Puhr ND, Lancette GA, Bodager DW, Toth BL, Genese CA, Highsmith AK, Pilot KE, Finelli L, Swerdlow DL.** 1998. Outbreak of *Salmonella* serotype Hartford infections associated with unpasteurized orange juice. JAMA **280:**1504-1509.
130. **Byrne L, Fisher I, Peters T, Mather A, Thomson N, Rosner B, Bernard H, McKeown P, Cormican M, Cowden J, Aiyedun V, Lane C.** 2014. A multi-country outbreak of *Salmonella* Newport gastroenteritis in Europe associated with watermelon from Brazil, confirmed by whole genome sequencing: October 2011 to January 2012. Euro Surveill **19:**6-13.
131. **Bayer C, Bernard H, Prager R, Rabsch W, Hiller P, Malorny B, Pfefferkorn B, Frank C, de Jong A, Friesema I, Stark K, Rosner B.** 2014. An outbreak of *Salmonella* Newport associated with mung bean sprouts in Germany and the Netherlands, October to November 2011. Euro Surveill **19:**20665.
132. **Schneider JL, White PL, Weiss J, Norton D, Lidgard J, Gould LH, Yee B, Vugia DJ, Mohle-Boetani J.** 2011. Multistate outbreak of multidrug-resistant *Salmonella* Newport infections associated with ground beef, October to December 2007. J Food Prot **74:**1315-1319.
133. **Cao G, Meng J, Strain E, Stones R, Pettengill J, Zhao S, McDermott P, Brown E, Allard M.** 2013. Phylogenetics and differentiation of *Salmonella* Newport lineages by whole genome sequencing. PLoS One **8:**e55687.
134. **Cao G, Allard MW, Hoffmann M, Monday SR, Muruvanda T, Luo Y, Payne J, Rump L, Meng K, Zhao S, McDermott PF, Brown EW, Meng J.** 2015. Complete Sequences of Six IncA/C Plasmids of Multidrug-Resistant *Salmonella enterica* subsp. *enterica* Serotype Newport. Genome Announcements **3:**e00027-00015.
135. **Chiu CH, Tang P, Chu C, Hu S, Bao Q, Yu J, Chou YY, Wang HS, Lee YS.** 2005. The genome sequence of *Salmonella enterica* serovar Choleraesuis, a highly invasive and resistant zoonotic pathogen. Nucleic Acids Res **33:**1690-1698.
136. **Kingsley RA, Msefula CL, Thomson NR, Kariuki S, Holt KE, Gordon MA, Harris D, Clarke L, Whitehead S, Sangal V, Marsh K, Achtman M, Molyneux ME, Cormican M, Parkhill J, MacLennan CA, Heyderman RS, Dougan G.** 2009. Epidemic multiple drug resistant *Salmonella* Typhimurium causing invasive disease in sub-Saharan Africa have a distinct genotype. Genome Res **19:**2279-2287.
137. **Werber D, Dreesman J, Feil F, van Treeck U, Fell G, Ethelberg S, Hauri AM, Roggentin P, Prager R, Fisher IS, Behnke SC, Bartelt E, Weise E, Ellis A, Siitonen A, Andersson Y, Tschape H, Kramer MH, Ammon A.** 2005. International outbreak of *Salmonella* Oranienburg due to German chocolate. BMC Infect Dis **5:**7.
138. **Kumao T, Ba-Thein W, Hayashi H.** 2002. Molecular subtyping methods for detection of *Salmonella enterica* serovar Oranienburg outbreaks. J Clin Microbiol **40:**2057-2061.
139. **Cummings KJ, Rodriguez-Rivera LD, Mitchell KJ, Hoelzer K, Wiedmann M,**

McDonough PL, Altier C, Warnick LD, Perkins GA. 2014. *Salmonella enterica* serovar Oranienburg outbreak in a veterinary medical teaching hospital with evidence of nosocomial and on-farm transmission. Vector Borne Zoonotic Dis **14:**496-502.

140. **den Bakker HC, Switt AI, Cummings CA, Hoelzer K, Degoricija L, Rodriguez-Rivera LD, Wright EM, Fang R, Davis M, Root T, Schoonmaker-Bopp D, Musser KA, Villamil E, Waechter H, Kornstein L, Furtado MR, Wiedmann M.** 2011. A whole-genome single nucleotide polymorphism-based approach to trace and identify outbreaks linked to a common *Salmonella enterica* subsp. *enterica* serovar Montevideo pulsed-field gel electrophoresis type. Appl Environ Microbiol **77:**8648-8655.
141. **Allard MW, Luo Y, Strain E, Pettengill J, Timme R, Wang C, Li C, Keys CE, Zheng J, Stones R, Wilson MR, Musser SM, Brown EW.** 2013. On the evolutionary history, population genetics and diversity among isolates of *Salmonella* Enteritidis PFGE pattern JEGX01.0004. PLoS One **8:**e55254.
142. **Barton Behravesh C, Mody RK, Jungk J, Gaul L, Redd JT, Chen S, Cosgrove S, Hedican E, Sweat D, Chavez-Hauser L, Snow SL, Hanson H, Nguyen TA, Sodha SV, Boore AL, Russo E, Mikoleit M, Theobald L, Gerner-Smidt P, Hoekstra RM, Angulo FJ, Swerdlow DL, Tauxe RV, Griffin PM, Williams IT.** 2011. 2008 outbreak of *Salmonella* Saintpaul infections associated with raw produce. N Engl J Med **364:**918-927.
143. **Safranek T, Leschinsky D, Keyser A, O'Keefe A, Timmons T, Holmes S, Garvey A, Von Stein D, Harris M, Quinlisk P, Bidol SA, Sheline KD, Collins JM, Vorhees R, Stella J, Ostroff S, Marriott C, Sandt CH, Chmieleski W, Lando J, Saathof-Huber L, Anderson S, Hedican E, Meyer S, Smith K, Miller B, Rigdon C, Salehi E, Kightlinger L, Schaefer L, Hepper C, Wilson S.** 2009. Outbreak of *Salmonella* serotype Saintpaul infections associated with eating alfalfa sprouts - United States, 2009. MMWR Morb Mortal Wkly Rep **58:**500-503.
144. **Taylor E, Kastner J, Renter D.** 2010. Challenges involved in the *Salmonella* Saintpaul outbreak and lessons learned. J Public Health Manag Pract **16:**221-231.
145. **Olson ND, Lund SP, Colman RE, Foster JT, Sahl JW, Schupp JM, Keim P, Morrow JB, Salit ML, Zook JM.** 2015. Best practices for evaluating single nucleotide variant calling methods for microbial genomics. Front Genet **6:**235.
146. **Leekitcharoenphon P, Nielsen EM, Kaas RS, Lund O, Aarestrup FM.** 2014. Evaluation of whole genome sequencing for outbreak detection of *Salmonella enterica*. PLoS One **9:**e87991.
147. **Sheppard SK, Jolley KA, Maiden MC.** 2012. A Gene-By-Gene Approach to Bacterial Population Genomics: Whole Genome MLST of *Campylobacter*. Genes **3:**261-277.
148. **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.
149. **Ashton P, Nair S, Peters T, Tewolde R, Day M, Doumith M, Green J, Jenkins C, Underwood A, Arnold C, de Pinna E, Dallman T, Grant K.** 2015. Revolutionizing Public Health Reference Microbiology using Whole Genome Sequencing: *Salmonella* as an exemplar, Nov 29, 2015. bioRxiv preprint. doi:10.1101/033225.
150. **Carleton H, Gerner-Smidt P.** 2016. Whole-Genome Sequencing Is Taking over Foodborne Disease Surveillance. Microbe **11:**311-317.
151. **Bekal S, Berry C, Reimer AR, Van Domselaar G, Beaudry G, Fournier E, Doualla-**

Bell F, Levac E, Gaulin C, Ramsay D, Huot C, Walker M, Sieffert C, Tremblay C. 2016. Usefulness of high-quality core genome single-nucleotide variant analysis for subtyping the highly clonal and the most prevalent *Salmonella enterica* serovar Heidelberg clone in the context of outbreak investigations. J Clin Microbiol **54:**289-295.

Appendix

Supplementary Table A: Outline of the serovars included in each panel, their number of representatives, and reason for selection.

Panel Number	Serovar Name	Antigenic Formula	Number of Isolates	Reason for Selection
Panel One	**4,[5],12:i:-**	4,[5],12:i:-	20	Top 20
	4,[5],12:b:-	4,[5],12:b:-	20	
	Agona	4,[5],12:f:g:s:[1,2]	20	
	Braenderup	6,7,14:e,h:e,n,z15	20	
	Enteritids	1,9,12:g,m:-	20	
	Hadar	6,8:z10:e,n,x	20	
	Heidelberg	4,[5],12:r:1,2	20	
	Infantis	6,7,14:r:1,5	20	
	Javiana	1,9,12:l,z28:1,5	20	
	Montevideo	6,7:14,[54]:g,m,[p],s:[1,2,7]	20	
	Muenchen	6,8:d:1,2	20	
	Newport	6,8,20:e,h:1,2	20	
	Oranienburg	6,7,14:m,t:[z57]	20	
	Paratyphi A	1,2,12:a:[1,5]	20	
	Paratyphi B var. Java	4,5,12:b:1,2 dT+	20	
	Saintpaul	4,[5],12:e,h:1,2	20	
	Stanley	4,[5],12,[27]:d:1,2	20	
	Thompson	6,7,14:k:1,5	20	
	Typhi	9,12[Vi]:d:-	20	
	Typhimurium	1,4,[5],12:i:-	20	
Panel Two	**Corvallis**	8,20:z4,z23:-	5	Travel Associated
	Cerro	18:z4,z23:-	5	Increased Invasive
	Dublin	9,12:g,p:-	4	
	Panama	9,12:l,v:1,5	5	
	Sandiego	4,5,12:e,h:e,n,z15	5	
	Schwarzengrund	4,12,27:d:1,7	5	
	Carrau	6,14:y:1,7	4	Difficult to Differentiate
	Kouka	1,3,19:g,m,t:-	1	
	Lattenkamp	45:z35:1,5	2	
	Madelia	6,14,25:y:1,7	2	
	Paratyphi B	4,[5],12:b:1,2 dT-	5	
	Senftenberg	1,3,19:g,s,t:-	2	
		ssp II 9,46:l,w:e,n,x	1	Non ssp I serovars
		ssp II 42:r:-	1	
		ssp II 47:b:1,5	1	
		ssp II 58:l,z13,z28:z6	2	
		ssp IIIa 41:z4,z23:-	2	

Panel Two		ssp IIIa 44:z4,z23,z32:-	1	Non ssp I serovars
		ssp IIIa 48:g,z51:-	1	
		ssp IIIa 48:z4,z24:-	1	
		ssp IIIb 11:k:z53	1	
		ssp IIIb 50:k:z	1	
		ssp IIIb 50:l,v:z35	1	
		ssp IIIb 50:z:z52	1	
		ssp IIIb 61:k:1,5,7	1	
		ssp IV 18:z36,z38:-	1	
		ssp IV 44:z4,z23:-	1	
		ssp IV 48:g,z51:-	1	
		ssp IV 48:z4,z32:-	1	
		ssp IV 50:g,z51:-	1	
	Untypeable	ssp I	8	Untypeable by traditional serotyping
		ssp IIIb	1	
		ORough:e,h:-	1	
		ORough:e,h:1,6	1	
		ORough:f,g,s:-	1	
		ORough:g,m:-	3	
		ORough:i:-	1	
		ORough:r:-	1	
		ORough:r:1,2	1	
		ORough:r:1,5	1	
		ORough:z:1,w	1	
		ORough:z10:e,n,x	1	
		ssp II ORough:c:z6	1	
		ssp IIIa ORough:g,z51:-	1	
		ssp IIIb ORough:z10:e,n,x,z15	1	
		ssp IV ORough:g:z51:-	1	
Panel Three	**Aberdeen**	11:i:1,2	1	Serovars of global importance and representatives of the majority of antigenic determinates
	Abony	1,4,[5],12,[27]:b:e,n,x	3	
	Adelaide	35:f,g:-	1	
	Agama	4,12:i:1,6	1	
	Agona	1,4,[5],12:f,g,s:[1,2]	3	
	Alachua	35:z4,z23:-	1	
	Albany	8,20:z4,z24:-	3	
	Altona	8,20:r,[i]:z6	1	
	Amsterdam	3,{10},{15},{15,34}:g,m,s:-	3	
	Anatum	3,{10},{15},{15,34}:e,h:1,6	3	
	Apapa	45:m,t:-	1	
	Banana	1,4,[5],12:m,t:[1,5]	1	
	Bareilly	6,7,14:y:1,5	1	
	Berlin	17:d:1,5	1	
	Berta	1,9,12:[f],g,[t]:-	1	

Panel Three	**Blegdam**	9,12:g,m,q:-	1	Serovars of global importance and representatives of the majority of antigenic determinates
	Blockley	6,8:k:1,5	3	
	Borreze	54:f,g,s:-	1	
	Bovismorbificans	6,8,20:r,[i]:1,5	3	
	Braenderup	6,7,14:e,h:e,n,z15	3	
	Brandenburg	4,[5],12:l,v:e,n,z15	3	
	Bredeney	1,4,12,27:l,v:1,7	3	
	California	4,12:g,m,t:[z67]	1	
	Carrau	6,14,[24]:y:1,7	1	
	Cerro	6,14,18:z4,z23:[1,5]	3	
	Chester	1,4,[5],12:e,h:e,n,x	3	
	Choleraesuis	6,7:c:1,5	6	
	Coeln	1,4,[5],12:y:1,2	1	
	Concord	6,7:l,v:1,2	1	
	Corvallis	8,20:z4,z23:[z6]	3	
	Crossness	67:r:1,2	1	
	Derby	1,4,[5],12:f,g:[1,2]	3	
	Dublin	1,9,12[Vi]:g,p:-	3	
	Entertidis	1,9,12:g,m:-	22	
	Gallinarum	1,9,12:-:-	2	
	Give	3,{10}{15}{15,34}:l,v:1,7	3	
	Godesberg	30:g,m,[t]:-	1	
	Goldcoast	6,8:r:l,w	1	
	Grumpensis	1,13,23:d:1,7	1	
	Hadar	6,8:z10:e,n,x	7	
	Havana	1,13,23:f,g,[s]:-	1	
	Heidelberg	1,4,[5],12:r:1,2	8	
	Hvittingfoss	16:b:e,n,x	1	
	4,[5],12:i:-	1,4,[5],12:i:-	1	
	Idikan	1,13,23:i:1,5	1	
	ssp II		3	
	ssp IIIa		3	
	ssp IIIb		6	
	ssp IV		2	
	S. bongori		2	
	Indiana	1,4,12:z:1,9	3	
	Infantis	6,7,14:r:1,5	7	
	Inverness	38:k:1,6	1	
	Isangi	6,7,14:d:1,5	1	
	Jangwani	17:a:1,5	1	
	Javiana	1,9,12:l,z28:1,5	3	
	Johannesburg	1,40:b:e,n,x	1	
	Karamoja	1,40:z41:1,2	1	
	Kedougou	1,13,23:i:l,w	3	
	Kentucky	8,20:i:z6	8	

Panel Three	**Kiambu**	1,4,12:z:1,5	3	Serovars of global importance and representatives of the majority of antigenic determinates
	Kiel	1,2,12:g,p:-	1	
	Kimuenza	1,4,12,27:l,v:e,n,x	1	
	Kisarawe	11:k:e,n,x,[z15]	1	
	Kottbus	6,8:e,h:1,5	3	
	Krefeld	1,3,19:y:l,w	1	
	Lexington	3,{10}{15}{15,34}:z10:1,5	1	
	Lille	6,7,14:z38:-	1	
	Litchfield	6,8:l,v:1,2	1	
	Liverpool	1,3,19:d:e,n,z15	1	
	Lingstone	6,7,14:d:l,w	3	
	Llandorff	1,3,19:z29:[z6]	1	
	London	3,{10}{15}:l,v:1,6	3	
	Macclesfield	9,46:g,m,s:1,2,7	1	
	Manchester	6,8:l,v:1,7	1	
	Manhattan	6,8:d:1,5	1	
	Matadi	17:k:e,n,x	1	
	Mbandaka	6,7,14:z10:e,n,z15	3	
	Meleagridis	3,{10}{15}{15,34}:e,h:l,w	1	
	Mikawasima	6,7,14:y:e,n,z15	1	
	Milwaukee	43:f,g,[t]:-	1	
	Minnesota	21:b:e,n,x	1	
	Mississippi	1,13,23:d:1,5	3	
	Monschaui	35:m,t:-	1	
	Montevideo	6,7,14,[54]:g,m,[p],s:[1,2,7]	7	
	Moscow	1,9,12:g,q:-	1	
	Muenchen	6,8:d:1,2	3	
	Muenster	3,{10}{15}{15,34}:e,h:1,5	3	
	Naestved	1,9,12:g,p,s:-	1	
	Newport	6,8,20:e,h:1,2	7	
	Nitra	2,12z:g,m:-	1	
	Ohio	6,7,14:b:l,w	1	
	Oranienburg	6,7,14:m,t:[z57]	3	
	Orion	3,{10}{15}{15,34}:y:1,5	3	
	Oslo	6,7,14:a:e,n,x	1	
	Ouakam	9,46:z29:-	1	
	Panama	1,9,12:l,v:1,5	3	
	Paratyphi A	1,2,12:a:[1,5]	3	
	Paratyphi B var. Java	1,4,[5],12:b:1,2	6	
	Pomona	28:y:1,7	1	
	Poona	1,13,22:z:1,6	1	
	Pullorum	1,9,12:-:-	1	
	Quebec	44:c:e,n,z15	1	
	Reading	1,4,[5],12:e,h:1,5	1	

Panel Three	**Rissen**	6,7,14:f,g:-	3	Serovars of global importance and representatives of the majority of antigenic determinates
	Rostock	1,9,12:g,p,u:-	1	
	Untypeable		1	
	Rubislaw	11:r:e,n,x	1	
	Saintpaul	1,4,[5],12:e,h:1,2	3	
	Sandiego	1,4,[5],12:e,h:e,n,z15	3	
	Schwarzengrund	1,4,12,27:d:1,7	3	
	Senftenberg	1,3,19:g,[s],t:-	3	
	Stanely	1,4,[5],12,[27]:d:1,2	4	
	Stanleyville	1,4,[5],12,[27]:z4,z23:[1,2]	3	
	Telelkebir	13,23:d:e,n,z15	1	
	Tennessee	6,7,14:z29:[1,2,7]	3	
	Thompson	6,7,14:k:1,5	3	
	Treforest	1,51:z:1,6	1	
	Typhi	9,12[Vi]:d:-	3	
	Typhimurium	1,4,[5],12:i:1,2	19	
	Uganda	3,{10}{15}:1,z13:1,5	1	
	Urbana	30:b:e,n,x	1	
	Utrecht	52:d:1,5	1	
	Virchow	6,7,14:r:1,2	7	
	Wandsworth	39:b:1,2	1	
	Waycross	41:z4,z23:-	1	
	Weltevreden	3,{10}{15}:r:z6	3	
	Weslaco	42:z36:-	1	
	Worthington	1,13,23:z:l,w	1	
	Yoruba	16:c:l,w	1	
	Zanzibar	3,{10}{15}:k:1,5	1	

Supplementary Table B: Performance of three *in silico* methods for *Salmonella* serovar prediction, SISTR, SeqSero, and 7-gene MLST, compared to traditional serotyping for 813 *Salmonella* isolates.

Serovar Groups	Total Tested	Full			Partial			Genotypic			Incorrect		
		SISTR	SeqSero	MLST	SISTR	SeqSero	MLST	SISTR	SeqSero	MLST	SISTR	SeqSero	MLST
Agona	**23**	22	22	22	0	0	1	0	0	0	1	1	0
Braenderup	**23**	23	17	23	0	0	0	0	0	0	0	6	0
Carrau	**5**	5	0	0	0	5	0	0	0	0	0	0	5
Cerro	**8**	8	0	7	0	5	0	0	0	0	0	3	1
Corvallis	**8**	8	0	8	0	8	0	0	0	0	0	0	0
Dublin	**7**	7	4	4	0	0	0	0	0	0	0	3	3
Enteritidis	**42**	40	34	40	1	2	1	0	0	0	1	6	1
Hadar	**27**	26	0	26	0	26	0	0	0	0	1	1	1
Heidelberg	**28**	28	27	28	0	1	0	0	0	0	0	0	0
I 4,[5],12:b:-	**20**	11	0	11	0	14	0	6	6	6	3	0	3
I 4,[5],12:i:-	**21**	18	0	0	0	20	0	1	1	1	2	0	20
Infantis	**27**	27	23	27	0	0	0	0	0	0	0	4	0
Javiana	**23**	23	0	23	0	20	0	0	0	0	0	3	0
Kouka	**1**	1	0	0	0	0	1	0	0	0	0	1	0
Lattenkamp	**2**	2	2	0	0	0	2	0	0	0	0	0	0
Madelia	**2**	2	0	0	0	2	0	0	0	0	0	0	2
Montevideo	**27**	26	21	25	0	0	1	0	0	0	1	6	1
Muenchen	**23**	23	0	23	0	23	0	0	0	0	0	0	0
Newport	**27**	26	26	26	0	0	0	0	0	0	1	1	1
Oranienburg	**23**	22	0	20	0	18	3	0	0	0	1	5	0
Panama	**8**	7	0	8	0	7	0	0	0	0	1	1	0
Paratyphi A	**23**	23	22	23	0	0	0	0	0	0	0	1	0
Paratyphi B	**5**	5	0	5	0	5	0	0	0	0	0	0	0
Paratyphi B var. Java	**26**	24	0	6	0	25	0	0	0	0	2	1	20
Rough O non ssp I	**5**	0	0	0	0	0	0	5	5	5	0	0	0
Rough O ssp I	**21**	0	0	0	0	0	0	21	21	21	0	0	0
Saintpaul	**23**	23	23	21	0	0	2	0	0	0	0	0	0
Sandiego	**8**	8	8	0	0	0	0	0	0	0	0	0	8
Schwarzengrund	**8**	8	8	8	0	0	0	0	0	0	0	0	0
Senftenberg	**7**	7	0	7	0	6	0	0	0	0	0	1	0
ssp II	**8**	6	1	0	0	4	8	0	0	0	2	3	0

ssp IIIa	**8**	7	3	1	0	4	7	0	0	0	1	1	0
ssp IIIb	**11**	9	2	0	0	5	11	0	0	0	1	4	0
ssp IV	**7**	7	2	0	0	5	1	0	0	0	0	0	6
Stanley	**24**	24	24	24	0	0	0	0	0	0	0	0	0
Thompson	**23**	23	20	23	0	1	0	0	0	0	0	2	0
Typhi	**23**	22	18	23	0	0	0	0	0	0	1	5	0
Typhimurium	**39**	39	38	39	0	0	0	0	0	0	0	1	0
Non-Target	**169**	139	95	132	8	38	14	0	0	0	19	36	23
Totals	**813**	**729**	**440**	**633**	**9**	**244**	**52**	**33**	**33**	**33**	**38**	**96**	**95**