A COMPARATIVE GENOMIC FRAMEWORK FOR THE *IN SILICO* DESIGN AND ASSESSMENT OF MOLECULAR TYPING METHODS USING WHOLE-GENOME SEQUENCE DATA WITH APPLICATION TO *LISTERIA MONOCYTOGENES*

PETER KRUCZKIEWICZ
Bachelor of Science, University of Victoria, 2010

A Thesis
Submitted to the School of Graduate Studies
of the University of Lethbridge
in Partial Fulfillment of the
Requirements for the Degree

MASTER OF SCIENCE

Department of Biological Sciences
University of Lethbridge
LETHBRIDGE, ALBERTA, CANADA

Dedicated to Mariola and Zbigniew Kruczkiewicz

Abstract

Although increased genome sequencing efforts have increased our understanding of genomic variability within many bacterial species, there has been limited application of this knowledge towards assessing current molecular typing methods and developing novel molecular typing methods. This thesis reports a novel *in silico* comparative genomic framework where the performance of typing methods is assessed on the basis of the discriminatory power of the method as well as the concordance of the method with a whole-genome phylogeny. Using this framework, we designed a comparative genomic fingerprinting (CGF) assay for *Listeria monocytogenes* through optimized molecular marker selection. *In silico* validation and assessment of the CGF assay against two other molecular typing methods for *L. monocytogenes* (multilocus sequence typing (MLST) and multiple virulence locus sequence typing (MVLST)) revealed that the CGF assay had better performance than these typing methods. Hence, optimized molecular marker selection can be used to produce highly discriminatory assays with high concordance to whole-genome phylogenies. The framework described in this thesis can be used to assess current molecular typing methods against whole-genome phylogenies and design the next generation of high-performance molecular typing methods from whole-genome sequence data.

Acknowledgements

I would like to thank my supervisors, Drs. James Thomas and Eduardo Taboada, without whose support this thesis would not have been possible. I am grateful for the time and energy they have invested in mentoring me and guiding me towards becoming and thinking like a scientist. I feel fortunate to have had supervisors that had my best interests at heart.

It was an honour to have Drs. Victor Gannon, Betty Golsteyn-Thomas and Brent Selinger on my thesis advisory committee. Their comments and suggestions lead me to much deliberation about the direction of the project. I would like to thank them for their guidance and support.

I would like to thank Dr. Dele Ogunremi for coming all the way from Ottawa to be the external examiner during my thesis defense. Thank you for the many comments and insights, and for the suggestions of potential future directions for this research. I would also like to thank Dr. Roy Golsteyn for chairing my thesis defense.

I would like to thank my colleagues at the Public Health Agency of Canada for their support and advice: Steven Mutschall and Dillon Barker for all of the wonderful suggestions for data analysis, software program features and algorithms; Cody Buchanan for the suggestions in improving various software programs I had developed; Cassandra Jokinen for all of the very useful and helpful advice for doing a masters; Ben Hetman for introducing me to Settlers of Catan; Chad Laing for the exchange of ideas and interesting and relevant papers; Susanna Whelpley for being so helpful and for dealing with HR on my behalf.

I would like to thank our collaborators at the National Microbiology Laboratory (Dr. Matt Gilmour and his *Listeria* genomics team, Dr. Morag Graham and the DNA core team, Dr. Gary Van Domselaar and his bioinformatics team) for providing access to *Listeria monocytogenes* whole-genome sequence data.

I would like to thank the Public Health Agency of Canada and the University of Lethbridge for their financial support, and the Government of Canada's Genomics Research and Development Initiative for partially funding this research.

Contents

List of Tables

List of Figures

Abbreviations

AFLP	**A**mplified **F**ragment **L**ength **P**olymorphism
BLAST	**B**asic **L**ocal **A**lignment **S**earch **T**ool
bp	**b**ase **p**air
CDS	**C**oding **D**NA **S**equence
CGF	**C**omparative **G**enomic **F**ingerprinting
CGH	**C**omparative **G**enomic **H**ybridization
CI	**C**onfidence **I**nterval
EC	**E**pidemic **C**lone
FTP	**F**ile **T**ransfer **P**rotocol
HGT	**H**orizontal **G**ene **T**ransfer
LVPC	**V**ariably-**P**resented **G**ene **C**lusters
M-CGH	**M**icroarray **C**omparative **G**enomic **H**ybridization
MIST	**M**icrobial ***In Silico*** **T**yper
MLEE	**M**ultilocus **E**nzyme **E**lectrophoresis
MLST	**M**ulti**l**ocus **S**equence **T**yping
MLVA	**M**ultilocus **V**ariable-number tandem repeats **A**nalysis
MSA	**M**ultiple **S**equence **A**lignment
MVLST	**M**ultiple **V**irulence **L**ocus **S**equence **T**yping
NCBI	**N**ational **C**enter for **B**iotechnology **I**nformation
NGS	**N**ext **G**eneration **S**equencing
NJ	**N**eighbor-**J**oining
NML	**N**ational **M**icrobiology **L**ab
ORF	**O**pen **R**eading **F**rame
PCR	**P**olymerase **C**hain **R**eaction
PFGE	**P**ulsed-**F**ield **G**el **E**lectrophoresis
PGC	**P**hylo**g**enomic **C**lusters
PHAC	**P**ublic **H**ealth **A**gency of **C**anada
PMCID	**P**ub**M**ed **C**entral **ID**
PMID	**P**ub**M**ed **ID**

RAPD	**R**andom **A**mplification Of **P**olymorphic **D**NA
RAST	**R**apid **A**nnotation using **S**ubsystems **T**echnology
RFLP	**R**estriction **F**ragment **L**ength **P**olymorphism
RF	**R**obinson-**F**oulds
SNP	**S**ingle **N**ucleotide **P**olymorphism
SSM	**S**lipped **S**trand **M**ispairing
ST	**S**equence **T**ype
UPGMA	**U**nweighted **P**air **G**roup with **A**rithmetic **M**ean
VNTR	**V**ariable-**N**umber **T**andem **R**epeats
WGS	**W**hole **G**enome **S**equencing

Symbols

AR	Adjusted Rand coefficient
AW	Adjusted Wallace coefficient
D	Simpson's index of diversity
R	Rand coefficient
W	Wallace coefficient

Chapter 1

Literature Review

1.1 Epidemiology

Epidemics of infectious disease have been the scourge of humanity since the beginnings of civilization. Throughout human history, much of our understanding of the spread of disease was anchored in superstition and myth. It was not until the cholera epidemic of 1854 in London that physician John Snow was able to determine how cholera was spread through investigations and analyses that became the foundation for modern epidemiology (Snow, 1857; Vinten-Johansen, 2003). Although *Vibrio cholerae* was not identified as the microbe responsible for causing cholera by Robert Koch until 1884 (Howard-Jones, 1984), Snow was able to determine that cholera was being spread by fecally contaminated water, and not by "bad air", which was thought to be the main route of transmission at the time (Vinten-Johansen, 2003). By plotting cholera mortality data, which included time and location of death, on a map of London and using other anecdotal evidence, he was able to determine that a well near Broad Street was the common source of infection in the district of Soho in London (Vinten-Johansen, 2003). Cases of cholera were clustered around this particular well. Anyone that drank from the well soon became afflicted. By presenting this evidence to the district officials, he was able to have the handle for the well removed effectively shutting down the well and ending the cholera epidemic in this district (Vinten-Johansen, 2003). Although not recognized for his contributions to medicine and epidemiology during his lifetime, he is often referred to as the Father of Epidemiology with many

of his descriptive and analytic epidemiologic approaches still used in epidemiology today (Merrill, 2010).

As more bacteria were discovered to cause disease in humans, public health efforts have been increased to assess potential routes of infection and formulate strategies for controlling and reducing the incidence of infectious disease caused by bacteria. Although in some cases traditional epidemiology may be sufficient to determine the source of an outbreak, other evidence in the form of bacterial strain-specific fingerprints are often used in conjunction with epidemiological data to unequivocally determine the source of an outbreak (Van Belkum et al., 2007). It has long been known that bacterial isolates from the same species exhibit variable levels of virulence, host-specificity and differences in terms of the modalities of disease. For example, within *Salmonella*, strains from one serotype such as Pullorum may only be found to cause disease within animals while strains from another serotype such as Typhi may only be associated with human clinical cases. In order to study the distribution and population structure of bacteria in clinical and environmental settings, various methods have been developed to differentiate strains or subtypes within a bacterial population (i.e., subtyping) through discrimination on the basis of phenotypic or genetic characteristics. Within the context of an epidemiological surveillance system, typing of bacterial isolates can be helpful in determining etiological agents, patterns of transmission and whether multiple sporadic cases of infection constitute a potential outbreak (Van Belkum et al., 2007).

1.2 Bacterial population diversity

Ever since we could observe and study bacteria, we have been trying to understand why certain strains of a bacterial species possess traits, such as the ability to cause disease, while other strains from the same species do not. In 1928, Griffith (1928) showed that bacteria can exchange genetic information through a process called transformation. Griffith found that if a killed virulent strain and a non-virulent strain of *Streptococcus pneumoniae* were injected into a mouse, living virulent bacteria could be recovered. The non-virulent strain had taken up and integrated the DNA of the killed virulent strain leading to a change in the phenotype. Bacteria have also been shown to exchange genetic information through recombination and conjugation. Bacteriophage integrating their DNA into the genomes of bacteria has helped explain the pathogenicity of certain strains

of bacteria such as *Escherichia coli* O157:H7 (Hayashi et al., 2001; Perna et al., 2001). These processes for increasing genetic diversity within bacterial species have likely evolved by natural selection (Smith et al., 1991). The new genetic information may repair host chromosome damage or confer a selective advantage, such as antimicrobial resistance (Smith et al., 1991). It is only recently with whole-genome sequences for multiple strains of many bacterial species that we are beginning to understand the extent of bacterial intra-species diversity.

Various typing methods have been developed to enable the characterization of bacterial isolates on the basis of intra-species diversity to aid in epidemiological surveillance and investigation. Even before the genetic basis for this diversity was fully understood, various methods have been used to differentiate isolates based on biochemical or serological profiles. As advances in molecular biology have enabled the direct interrogation of genetic variation, molecular typing methods have been developed offering greater discriminatory power, reproducibility and other advantages over phenotypic methods (Van Belkum et al., 2007). This has given rise to molecular epidemiology where molecular biology techniques are applied to epidemiologic problems (Foxman and Riley, 2001). Along with greatly benefiting epidemiological investigations and surveillance, advances in typing methods have also enabled the study of the population structure and population dynamics of bacterial species leading to an increased understanding of the epidemiology and evolution of bacterial populations (Smith et al., 1993).

1.3 Phenotypic subtyping

Before genetic differences between strains could be directly observed using molecular methods, phenotypic methods were used as a proxy for assessing genetic similarity between strains. Various phenotypic typing methods were developed for characterization of bacterial isolates and epidemiological surveillance. Some phenotypic typing methods have been more successfully applied to characterization of bacterial species than others. Phenotypic methods have provided historically and epidemiologically important insights into the population structure and genetic diversity of many bacterial species. Although most phenotypic methods have been replaced by molecular methods that can directly target genetic variation, some phenotypic subtyping methods are still in use today.

1.3.1 Biotyping

Biotyping is a typing method that differentiates bacterial isolates on the basis of biochemical characteristics that are known to vary within a given species (Van Belkum et al., 2007). Biotyping is usually simple and inexpensive with excellent typeability and data that are easy to score and interpret. Testing of large numbers of isolates can be performed easily. However, a large number of characteristics may need to be tested to provide sufficient discriminatory power (Van Belkum et al., 2007). In cases where a biotyping method is shown to be reproducible, it can be used as a library typing method where current isolates can be compared against previously characterized isolates (Van Belkum et al., 2007). Various biotyping arrays have been developed that utilize redox chemistry to enable measurement of biochemical reactions through colour readings (Bochner et al., 2001; Odumeru et al., 1999). These arrays have been useful in studying biochemical pathways and intra-species phenotypic variability (Baumler et al., 2011). However, the reproducibility of biotyping can be limited depending on the organism and phenotypic trait that is being tested (Miller and Rhoden, 1991; Odumeru et al., 1999).

1.3.2 Serotyping

First described by Lancefield in 1933, serotyping is one of the oldest phenotypic methods for the characterization and classification of bacterial isolates. Conventional serotyping is able to discriminate bacterial isolates on the basis of agglutination reactions of antisera with cell surface antigens of bacteria.

Historically, serotyping has been very useful for the epidemiological surveillance and classification of many bacterial pathogens such as *Listeria monocytogenes*, *Salmonella enterica* and *E. coli* (Kauffmann, 1975; McLauchlin, 1990; Salmonella Subcommittee of the Nomenclature Committee of the International Society for Microbiology, 1934; Seeliger and Langer, 1989). Since the 1930s, serotyping has been one of the primary methods for the typing of *Salmonella* (Kauffmann, 1975; Salmonella Subcommittee of the Nomenclature Committee of the International Society for Microbiology, 1934). However, in other organisms, serotyping has shown conflicting results with molecular methods (Aarts et al., 1995). Additionally, cross-reactions of antigens with other bacterial species may produce non-specific results (Doumith et al., 2004a; Palumbo et al., 2003;

Seeliger and Langer, 1989). Furthermore, some isolates may not be typeable by serotyping due to the absence of antigens or antisera for the antigens (Aeschbacher and Piffaretti, 1989).

Although serotyping can be very useful for subtyping certain bacterial species, conventional serotyping through slide agglutination using reactions with antisera is labour intensive, time-consuming and costly due to the expensive antisera required (Doumith et al., 2004a; Palumbo et al., 2003; Seeliger and Langer, 1989). To address some of the issues with conventional serotyping, serotyping enzyme-linked immunosorbent assay (ELISA) have been developed (McConnell et al., 1985; Palumbo et al., 2003). Molecular serotyping assays have also been developed to provide rapid, inexpensive and reliable serotyping of isolates (Doumith et al., 2004a; Franklin et al., 2011; Kérouanton et al., 2010; Scaria et al., 2008; Yoshida et al., 2007). However, the coverage of these molecular serotyping methods may only distinguish a limited number of serotypes and might require other subtyping methods in epidemiological investigations (Franklin et al., 2011; Kérouanton et al., 2010; Yoshida et al., 2007). Additionally, despite serotyping being an excellent characterization method for some bacteria, there is often little correlation between the somatic and flagellar antigens and genomic characteristics; therefore, molecular typing methods have been suggested as an alternative (Achtman et al., 2012).

1.3.3 Phage typing

Phage typing is the characterization of bacterial strains based on susceptibility to a standard set of bacteriophage. It has been very useful for the subtyping of *L. monocytogenes* and *Staphylococcus aureus* (Audurier and Martin, 1989; Wentworth, 1963). However, isolates are often not typeable by phage typing, which can be a problem with data analysis since all untypeable isolates will have the same data regardless of true relatedness (Aeschbacher and Piffaretti, 1989). Although phage typing may remain a useful tool for typing various pathogens into the near future, molecular typing methods providing greater resolution and accuracy will continue to replace most phenotypic methods including phage typing (Baggesen et al., 2010).

1.4 Molecular subtyping

There has been a shift from phenotypic subtyping towards molecular subtyping for many organisms as molecular methods have been developed allowing direct investigation of genetic variation between bacterial isolates. The emergence of molecular methods started with the discovery of restriction enzymes, the development of DNA sequencing and polymerase chain reaction (PCR) leading to the development of a number of techniques aimed at identifying genetic variation within bacterial isolates. Most molecular or genotypic typing methods offer higher throughput, lower workload, and greater reproducibility and subtyping resolution compared to phenotypic typing methods (Van Belkum et al., 2007). From an epidemiological and evolutionary perspective, molecular methods allow more accurate estimation of genetic similarity between isolates and more accurate inference of population structure.

1.4.1 Pulsed-field gel electrophoresis

Pulsed-field gel electrophoresis (PFGE) is one of the most commonly used molecular typing methods. PFGE involves the digestion of genomes using rare-cutting endonucleases and visualization of the resulting bands using gel electrophoresis in a unit capable of generating electrical fields that alternate in direction, which is necessary in order to resolve large DNA fragments (Finney, 2001). PFGE was one of the first methods that enabled investigation of genomic variability on a whole-genome level between bacterial isolates of the same species. Since insertions, deletions or mutations that create or remove a specific restriction site can be detected between the genomes of bacterial isolates, PFGE can be used for the characterization of strains and the indirect determination of evolutionary distances between isolates (Goering, 2010). Hence, PFGE was one of the first tools used for genomics studies.

There are many issues with performing PFGE for the characterization of bacterial isolates. PFGE requires the isolation of intact genomic DNA for comparable results between different samples; therefore, a specialized DNA isolation method is required (Goering, 2010). In addition to selection of appropriate DNA isolation methods, selection of restriction enzymes, electrophoresis conditions and many other variables influence how the DNA fragmentation banding patterns appear in the final gel electrophoresis image (Boerlin et al., 1996; Goering, 2010; Yde and Genicot, 2004).

Additionally, in order to achieve sufficient resolution for subtyping certain organisms, PGFE analysis with at least two restriction enzymes may be required (Boerlin et al., 1996; Yde and Genicot, 2004). Although it is possible to visually inspect PFGE gel images for differences between isolates, specialized and costly commercial computer-assisted analysis software is required for rigorous analysis of the PFGE gel images (Goering, 2010). In order to facilitate sharing of PFGE results between labs, standardized international protocols have been developed for many food-borne pathogens, such as *L. monocytogenes*, *E. coli*, *Salmonella* and *Vibrio parahaemolyticus*, through the PulseNet surveillance system (Graves and Swaminathan, 2001; Pagotto et al., 2006; Parsons et al., 2007; Ribot et al., 2006; Swaminathan et al., 2001; Swaminathan and Gerner-Smidt, 2007). Due to its simplicity and broad applicability, it is still used as a gold-standard typing method for many bacterial pathogens though it is a time-consuming and labour-intensive method (Goering, 2010).

1.4.2 PCR-based molecular subtyping

The invention of PCR lead to the development of various molecular typing methods potentially offering many advantages in terms of labour, cost, throughput and reproducibility (Killgore et al., 2008).

1.4.2.1 PCR-RFLP

Restriction fragment length polymorphism (RFLP) uses frequent cutting restriction enzymes to generate hundreds of small fragments from chromosomal DNA. With PCR-RFLP, genetic regions are amplified by PCR, digested by restriction enzymes, followed by separation of the resulting fragments by electrophoresis (Van Belkum et al., 2007). The resulting fingerprints can be compared. Fingerprints will vary due to mutations occuring within restriction sites leading to differences in the overall restriction fragment pattern. Various PCR-RFLP methods have been developed for the identification of bacterial species (Vaneechoutte et al., 1998) and the typing of bacterial species (Nachamkin et al., 1993; Sugimoto et al., 2011). Although, in some cases, PCR-RFLP may have less discriminatory power than other typing methods (Behringer et al., 2011; Sugimoto et al., 2011), like other PCR-based methods, it can be used to rapidly and easily type bacterial isolates (Sugimoto et al., 2011; Van Belkum et al., 2007).

1.4.2.2 Random amplified polymorphic DNA typing

Random amplified polymorphic DNA (RAPD) typing uses single primers of arbitrary sequence for low stringency PCR amplification to generate strain-specific DNA fragment banding patterns (Williams et al., 1990). Extensive genetic diversity was found between human clinical isolates of *Helicobacter pylori* using RAPD (Akopyanz et al., 1992). For the typing of *E. coli*, RAPD was found to have greater sensitivity than multilocus enzyme electrophoresis (Wang et al., 1993). Unlike PFGE, RAPD typing does not require intact chromosomes or large quantities of DNA, but it could theoretically be used to type any isolate just like PFGE. However, as with PFGE and amplified fragment length polymorphism typing (AFLP), RAPD typing markers are "anonymous" meaning it is unknown what part of the genome a particular band on a gel represents. Additionally, there are many issues with reproducibility because of the random priming and because PCR is performed under low stringency conditions.

1.4.2.3 Amplified fragment length polymorphism typing

AFLP typing is based on selective PCR amplification of genomic DNA restriction fragments using arbitrary primers and gel analysis of the amplified fragments (Vos et al., 1995). A major advantage of AFLP over other fragment-based typing methods is that data can be captured and digitized in an automated fashion using fluorescently labeled primers with an automatic DNA sequencing instrument (Van Belkum et al., 2007). AFLP has been used to successfully differentiate *Acinetobacter* species (Janssen et al., 1997) and outbreak and non-outbreak *Acinetobacter baumannii* isolates (Dijkshoorn et al., 1996). AFLP was found to produce typing data concordant with PFGE for *Pseudomonas aeruginosa* and *A. baumannii* (D'Agata et al., 2001). However, for vancomycin-resistant *Enterococcus faecium*, AFLP lacked the resolution of PFGE clustering epidemiologically-related and unrelated strains together (D'Agata et al., 2001). Although AFLP has been shown to produce comparable data to PFGE, in some cases it fails to provide sufficient resolution for epidemiological investigation.

1.4.2.4 Multilocus variable-number tandem-repeat analysis

Multilocus variable-number tandem-repeat (VNTR) analysis (MLVA) can be used to discriminate bacterial isolates on the number of observed repeats at regions of repetitive DNA (Murphy et al.,

2007). During DNA replication, repetitive DNA is often incorrectly copied due to slipped strand mispairing (SSM) resulting in the addition or deletion of repeats. Using PCR assays, it is possible to amplify repeat regions using primers flanking the region. The number of repeats found in the repeat region can be determined based on the size of the PCR product. Several MLVA schemes have been developed for subtyping *L. monocytogenes* targeting different VNTR loci (Lindstedt et al., 2008; Miya et al., 2008; Murphy et al., 2007; Sperry et al., 2008). MLVA has been shown to be a rapid, inexpensive, and discriminatory method for subtyping *L. monocytogenes* (Balandyt et al., 2011; Dass et al., 2010; Miya et al., 2008; Murphy et al., 2007) that may replace PFGE as the gold-standard method (Vergnaud and Pourcel, 2009). However, since similar VNTR profiles in unrelated strains may arise by convergent evolution and repeat regions often evolve rapidly, it may be difficult to infer accurate phylogenetic relationships between strains based on MLVA results, thus, hindering epidemiological and evolutionary analyses (Comas et al., 2009). Furthermore, genetic distances can be overestimated due to the high discriminatory power and the limited number of loci targeted by many MLVA schemes. This means that in the case of an outbreak, if two strains are identical by MLVA, it is likely that the strains are very similar (Van Belkum et al., 2007). Conversely, relationships inferred between isolates may not be concordant with clinical or epidemiologic information, and two strains very different by MLVA may actually be genetically quite similar (Xiao et al., 2011).

1.5 Population genetics and typing

The emergence of methods for assessing allele differences of genes within bacterial populations enabled population genetic analysis and the determination of genetic relationships between strains (Aeschbacher and Piffaretti, 1989; Smith et al., 1991, 1993). An important analysis in population genetics is determining linkage of alleles at different loci between bacterial isolates from different populations. Linkage equilibrium occurs in populations where alleles at different loci appear randomly exhibiting no association with alleles at other loci (Smith et al., 1991). In other words, there is a high level of genetic exchange between all members of a population. Highly recombinogenic bacterial species, such as *H. pylori* or *Neisseria meningitidis*, display linkage equilibrium and extensive genetic diversity (Van Belkum et al., 2007). On the other hand, linkage disequilibrium – the non-random association between alleles at different loci – can be observed in many bacterial species such as *Salmonella* or *E. coli* (Smith et al., 1993). Linkage disequilibrium

occurs when there is limited genetic exchange between members of a population. Genotypically distinct populations exhibiting linkage disequilibrium may be found to be geographically or ecologically separated or there may be biological barriers to genetic exchange (Smith et al., 1993). For example, *E. coli* possess a high level of genetic variation yet are naturally clonal (Smith et al., 1991). Recombination through conjugation is not common enough to extensively randomize the genome of *E. coli*, hence, only a small fraction of the possible clones are observed (Smith et al., 1991). There are also cases where genetically diverse species of bacteria, such as *N. meningitidis*, exhibit temporary disequilibrium leading to an epidemic population structure (Smith et al., 1993). Certain clones of the population will become dominant through clonal expansion, like in the case of an outbreak, and form complexes of related strains (Smith et al., 1993). This can result in an epidemic population structure where there is frequent recombination between all members of this population. While a highly clonal population with no recombination between members can be represented by a tree with distinct ancestors, an epidemic population can be represented as more of a network since there is swapping of genetic material between all members (Smith et al., 1993). Bacterial population genetics studies have lead to a greater understanding of the epidemiology of bacterial populations and the mechanisms that govern the exchange of genetic information within bacterial populations.

1.5.1 Multilocus enzyme electrophoresis

Although multilocus enzyme electrophoresis (MLEE) had been used extensively within eukaryotic genetics and evolutionary biology (Aeschbacher and Piffaretti, 1989), it was due to work by Selander et al. (1986) and others that the usefulness of MLEE in bacterial population genetics and characterization was recognized. With MLEE, isolates are characterized by relative electrophoretic mobilities of a set of housekeeping enzymes encoded by different alleles of the same gene (Selander et al., 1986). By assessing variation in the size and net electrostatic charge and, hence, the rate of migration of these housekeeping enzymes during electrophoresis, MLEE allows determination of allelic variation in the housekeeping genes associated with these enzymes (Aeschbacher and Piffaretti, 1989; Selander et al., 1986). Due to the enzymatic properties of these proteins, they can be visualized after electrophoresis as specific bands via reactions with certain substrates (Aeschbacher and Piffaretti, 1989; Selander et al., 1986).

Although MLEE is a phenotypic method and is neither rapid nor was it ever widely-used for bacterial typing (Van Belkum et al., 2007), it allowed consistent characterization of isolates for epidemic analyses and measurement of genetic distances between strains (Aeschbacher and Piffaretti, 1989). Furthermore, a significant advantage of MLEE compared to other methods at the time was that MLEE data allowed population genetics analysis through the study of the linkage of alleles at different loci within bacterial populations (Selander et al., 1987; Smith et al., 1993; Whittam et al., 1983).

Despite the importance of MLEE in population genetics, it was largely replaced by its molecular equivalent, multilocus sequence typing (MLST), which offers greater resolution, more rapid analysis and a high level of transportability of results due to the generation of DNA sequence data rather than gel images (Van Belkum et al., 2007).

1.5.2 Multilocus sequence typing

MLST is a DNA sequence-based typing method where bacterial isolates are differentiated based on sequence variation between alleles at typically 6 to 8 housekeeping genes (Maiden et al., 1998; Van Belkum et al., 2007). MLST is the molecular descendant of multilocus enzyme electrophoresis (MLEE), which was mostly used for population genetics rather than typing. Since DNA sequence data is the output of MLST, inter-laboratory data-sharing is simple and easy through the Internet. With publicly accessible online databases such as PubMLST (Jolley et al., 2004), MLST DNA sequence data and associated epidemiological information is hosted and shared between labs allowing for comprehensive epidemiological analyses (Chan et al., 2001; Jolley et al., 2004).

MLST can resolve a greater number of alleles per locus than MLEE thereby offering greater discriminative power than MLEE (Maiden et al., 1998), which it has largely replaced for population genetics analysis (Van Belkum et al., 2007). While MLEE can only detect variation in enzyme amino acid sequences resulting in a change in net charge, MLST can differentiate alleles that vary by a single-nucleotide polymorphism (SNP) even though the SNP might not result in a change at the amino acid level. Furthermore, MLST provides a less technical and more rapid approach to characterization of bacterial isolates, which has translated to its widespread usage for the molecular typing of bacterial organisms (Van Belkum et al., 2007). Hence, MLST replaced MLEE in most population genetics studies.

Algorithms such as the eBURST algorithm (Feil et al., 2004) and the globally optimized implementation of the eBURST (goeBURST) algorithm (Francisco et al., 2009) have allowed the sophisticated analysis of MLST data. Investigation of potential patterns of evolutionary descent and inference of the epidemic population structure of bacterial pathogens has been possible through application of the eBurst and goeBURST algorithms to vast databases of MLST typing data.

Although MLST has been very important in population genetics and molecular subtyping of many bacterial organisms, it is not an appropriate typing method for some bacteria as it is possible to infer incorrect phylogenetic relationships between strains using MLST data. In monomorphic bacterial species, such as *Mycobacterium tuberculosis*, MLST is unable to provide sufficient phylogenetic resolution for subtyping bacterial isolates (Comas et al., 2009; Musser et al., 2000; Sreevatsan et al., 1997). In highly recombinogenic bacteria, such as *Campylobacter jejuni*, strains with similar MLST profiles may actually have significant genomic differences (Taboada et al., 2008). Additionally, MLST is an expensive and time-consuming typing method where discriminative ability is dependent on the gene loci selected (Cai et al., 2002; Van Belkum et al., 2007). In some clonal bacteria, such as *L. monocytogenes*, targeting virulence genes in an MLST scheme may produce an assay with higher discriminative power than a regular MLST scheme targeting housekeeping genes (Knabel et al., 2012; Zhang et al., 2004). However, by selecting non-housekeeping genes, genes may be targeted that are under selective pressure, which can lead to the inference of inaccurate evolutionary relationships. With whole-genome sequencing costs continuously dropping, many laboratories may opt to produce sequence data for the entire genome of their bacterial strains rather than a select few gene loci.

1.6 Comparative genomics

1.6.1 Early whole-genome sequencing

In 1977, Sanger et al. described a DNA sequencing method using chain-terminating inhibitors of DNA polymerase. This sequencing method, termed Sanger sequencing, was the primary DNA sequencing method for sequencing of the human genome (Venter et al., 2001) and many other genomes prior to the advent of high-throughput next-generation sequencing (NGS) technologies.

Although Sanger sequencing was prohibitively expensive for many researchers, the sequencing of bacterial strains from various species allowed the development of high-resolution methods for investigation of bacterial intra-species genetic diversity.

Although evidence for extensive intra-species genetic diversity predates the era of genomics, it wasn't until comparative genomic analysis of two unrelated strains of *H. pylori* was performed that there was conclusive evidence for large-scale gene content differences between strains of the same species (Alm et al., 1999). Studies in *E.coli* showed evidence for large-scale genomic variability when the enterohemorrhagic *E.coli* O157:H7 EDL933 was compared to the non-pathogenic *E.coli* K-12 (Perna et al., 2001). Many genomic regions found within the genome of O157:H7 were not present within K-12. Some of these genomic regions present within O157:H7 were linked to virulence attributes (Perna et al., 2001). These studies and others have contributed to an emerging view of bacterial genomics in which there exists extensive genetic diversity between strains of the same species (Israel et al., 2001; Kato-Maeda et al., 2001; Taboada et al., 2004). This and mounting evidence that genetic variation between bacterial strains of the same species can lead to variation in host specificity and virulence (Merchant-Patel et al., 2008) fueled further exploration of intra-species gene content variability and led to the emergence of the field of bacterial comparative genomics.

1.6.2 Microarray-based comparative genomic hybridization

DNA microarray technology, which was originally used for gene expression studies (Schena et al., 1995), was adapted to study genome-wide variation (Lashkari et al., 1997). Microarray-based comparative genomic hybridization (M-CGH) was used to assess the absence or presence of thousands of genes in a single experiment. At a time when DNA sequencing was prohibitively expensive for comparative genomics studies of more than a few strains, M-CGH provided an effective approach to leveraging the data from a single sequenced strain towards more comprehensive investigation of the genome dynamics and the "phylogenomics" of a species (Taboada et al., 2007).

Comparative genomics studies using M-CGH lead to insights into the population structure and intra-species diversity of many bacterial species such as *C. jejuni*, *E. coli* and *L. monocytogenes* (Dorrell et al., 2001; Doumith et al., 2004b; Taboada et al., 2008; Zhang et al., 2003, 2007).

M-CGH studies of *E. coli* O157:H7 revealed differences in virulence-associated loci between the lineages (Zhang et al., 2007). In *C. jejuni*, extensive gene content differences were noted even between isolates that shared similar MLST profiles (Taboada et al., 2008). Comparative genomics studies of *L. monocytogenes* confirmed a highly clonal population structure and revealed serotype- and lineage-specific gene content differences (Call et al., 2003). Although the idea was explored, M-CGH was never widely adopted for high-resolution molecular subtyping of bacterial isolates (Lucchini et al., 2001) due to its high cost and low throughput (Van Belkum et al., 2007). Hence, M-CGH was mostly used for comparative genomics studies, the assessment of conventional typing methods and the screening of molecular markers including those associated with virulence factors (Chizhikov et al., 2001; Doumith et al., 2006; Taboada et al., 2008).

1.7 The pan-genome paradigm for bacteria

The first major sequence-based comparative genomic survey of a species published was on an ambitious project to sequence several strains of *Streptococcus agalactiae*, a human pathogen that is the major cause of septicemia in newborns (Tettelin et al., 2005). This effort provided evidence for the existence of intra-species "pan-genomes" – a pan-genome being the totality of genes observed within all members of a bacterial species (Medini et al., 2005; Tettelin et al., 2005). In the pan-genome model, a subset of genes representing the core genome is present in each member of the species (Tettelin et al., 2005). These are the "housekeeping" genes that perform vital functions for the organism. The remaining genes within the species' pan-genome represent the dispensable or accessory genome. Although these genes are not necessary for basic survival, they are thought to be required for niche adaptation and may also be associated with clinically important traits such as virulence, antimicrobial resistance or alternate metabolic pathways (Tettelin et al., 2005).

Pan-genomic analyses have shown that there is an extensive pool of genes within most bacterial species' pan-genomes that have yet to be identified and characterized (Tettelin et al., 2005). A new strain of *S. galactiae*, for example, could add more than 30 new genes to the gene pool for the species (Tettelin et al., 2005). Hence, it is necessary to sequence strains from various sources in order to understand the distribution of gene content across various subsets of the population (Medini et al., 2005; Tettelin et al., 2005). In a public health context, strain specific

accessory genes may be valuable indicators of virulence or pathogenic potential (Atherton, 1998; Duong and Konkel, 2009; Fouts et al., 2005) and may thus represent valuable targets for the identification of clinically important strains. The possibility of gene content variability within a bacterial species and its public health implications has lead to efforts to understand the overall microbial pan-genome as well as the pan-genomes of individual species (Alcaraz et al., 2010; Lukjancenko et al., 2010; Medini et al., 2005; Monot et al., 2009; Snipen et al., 2009; Tettelin et al., 2008; Xu et al., 2010).

Although gene absence/presence is an important source of genetic variability that can be an informative indicator of the potential virulence of a strain, other types of genetic variation exist that may be equally important. Another common type of genetic variation is single nucleotide polymorphisms (SNPs), which are insertions, substitutions or deletions of single nucleotides within a genome. SNPs can have important phenotypic implications. For instance, SNPs in the *inlA* gene of *L. monocytogenes*, responsible for host cell invasion, are correlated with varying levels of virulence due to the presence of premature stop codons (Van Stelten et al., 2010).

1.8 Next-generation sequencing and genomic epidemiology

The advent of high-throughput NGS has lead to a significant drop in sequencing costs with sequence data being generated at an increasingly rapid rate. With the introduction of pyrosequencing (Margulies et al., 2005) and Illumina sequencing (Bentley et al., 2008) less than a decade ago, whole-genome sequencing has become much more affordable. This has intensified efforts towards studying the pan-genomes of many bacterial species and, by extension, the intra-species genetic diversity within those species (Alcaraz et al., 2010; den Bakker et al., 2010; Deng et al., 2010; Lukjancenko et al., 2010). These and upcoming advances in sequencing technologies will continue to be a boon for comparative genomics of microbes.

High-throughput whole-genome sequencing (WGS) has rapidly become a viable option for the investigation of human pathogens in the context of outbreaks (Alexander et al., 2012; Gilmour et al., 2010; Rasko et al., 2011). Advances in WGS technologies and bioinformatic analyses necessary for the assembly and annotation of short nucleotide reads generated by shotgun WGS

have allowed WGS to emerge as a powerful tool for epidemiological investigations. Since WGS data can, in principle, differentiate between isolates differing by one nucleotide, WGS offers the highest level of discrimination for bacterial strain characterization and the inference of the most accurate phylogenetic relationships.

Epidemiological investigations into recent outbreaks have been greatly aided by WGS of outbreak strains (Alexander et al., 2012; Gardy et al., 2011; Rasko et al., 2011; Reimer et al., 2011; Rohde et al., 2011). Comparative genomics of whole-genome sequenced *V. cholerae* isolates associated with the 2010 Haiti cholera outbreak enabled the potential source of the outbreak strain to be determined whereas conventional subtyping methods, such as PFGE, lacked the resolution necessary for such conclusions (Reimer et al., 2011). With the 2011 *E. coli* O104:H4 outbreak in Germany, new WGS technologies and "crowd-sourced" analysis of the genomic data revealed that the outbreak strain belonged to an enteroaggregative *E. coli* lineage that had horizontally acquired virulence and antibiotic resistance genes (Rasko et al., 2011; Rohde et al., 2011). Using WGS, two travel-associated cases of *E. coli* O104:H4 were characterized with one case found to be related to the 2011 *E. coli* O104:H4 outbreak in Germany (Alexander et al., 2012). WGS aided an epidemiological investigation into a tuberculosis outbreak in a community in British Columbia, Canada by providing the resolution necessary to differentiate between *M. tuberculosis* isolates identical by conventional typing methods (Gardy et al., 2011). Eventually, WGS may be the sole method needed for molecular characterization of human bacterial pathogens during an outbreak, however, for epidemiological surveillance, WGS will likely remain economically infeasible for the foreseeable future.

1.9 Challenges with current molecular typing methods

Due to advances in NGS technologies, the throughput of sequencing entire genomes has increased exponentially while sequencing costs have dropped precipitously (Bennett, 2004; Margulies et al., 2005). This has lead to a democratization and acceleration of sequencing efforts where sequencing projects can be conducted by labs of any size resulting in a greater understanding and investigation of bacterial pan-genomes (Medini et al., 2005; Shendure and Ji, 2008). Concomitant with the development of new sequencing methods has been the continual development of bioinformatic approaches for the analysis of this data. Although new algorithms and approaches are continually

being developed for genome assembly, annotation and visualization (Pop and Salzberg, 2008), most typing methods currently used for the identification and characterization of human bacterial pathogens available to public health laboratories predate the genomic era. There has been limited application of comparative genomic approaches using WGS data towards development of diagnostic and molecular typing assays.

Although many current molecular typing methods may be well suited for certain types of epidemiological investigations, these methods are confounded by homoplasy, homologous recombination and horizontal gene transfer (HGT) leading to flawed estimates of genetic similarity between strains (Achtman and Wagner, 2008). Therefore, in order to infer the most accurate phylogenetic relationships between strains, WGS data is necessary to limit the influence of these confounders in the construction of a phylogeny (Pearson et al., 2009). When Comas et al. (2009) compared the phylogeny derived from 89 coding genes sequences from a global strain selection of *M. tuberculosis* to CRISPR- and VNTR-based typing results for the same strains, they found that although the CRISPR- and VNTR-based typing methods were highly discriminatory, the phylogenetic relationships inferred by these typing methods showed little statistical support and limited agreement with the coding gene sequence-derived phylogeny. Comas et al. (2009) concluded that selection of a combination of lineage-defining markers and highly discriminatory markers may be necessary to overcome the limitations of current typing methods for monomorphic microbes.

Using a combination of lineage-defining markers and highly discriminatory markers, Xiao et al. (2011) developed a molecular typing scheme for *V. parahaemolyticus* targeting VNTR regions and variably-presented gene clusters (LVPCs). The LVPC-based method clustered strains into broad groups that corresponded with distinct virulence profiles with clinical and epidemiological significance while the VNTR-based method provided greater discrimination producing finer groups. However, the VNTR-derived clusters did not correspond with distinct virulence profiles as with the LVPC-derived clusters. Therefore, Xiao et al. (2011) concluded that in combination targeting VNTR and LVPC loci could yield a high performance typing scheme for *V. parahaemolyticus*. Hence, for certain bacterial species, it may be necessary to use a combination of typing methodologies targeting different types of genetic variation to accurately characterize isolates (Clark et al., 2012).

Through comparison of DNA sequences, single-nucleotide polymorphisms (SNPs) can be identified for the differentiation of bacterial strains. Using phylogenetically informative SNP sites from

DNA sequence analysis of 65 *L. monocytogenes* isolates, Ducey et al. (2007) developed an assay for subtyping lineage I *L. monocytogenes*, which was later expanded for subtyping of all four lineages of *L. monocytogenes* (Ward et al., 2008). A SNP typing assay has also been developed for the rapid determination of epidemic clone (EC), serotype and lineage of *L. monocytogenes* isolates (Ward et al., 2010), and another SNP typing assay has been developed for the determination of premature stop codons in the *inl*A gene of *L. monocytogenes* for determination of isolate virulence (Van Stelten and Nightingale, 2008). However, there is a significant initial cost for SNP typing due to the specialized equipment required, which may prohibit adoption for routine surveillance. Furthermore, since SNP assays are retrospective in nature, novel lineages may not be detected since only previously observed SNPs may be targeted by the assay.

Van Belkum et al. (2007) define a good bacterial typing method as one that is inexpensive, rapid, epidemiologically concordant, discriminatory and reproducible. However, many conventional typing methods provide sub-par performance in one or more of these categories. Since many researchers are advocating the use of multiple typing methods for the characterization of bacterial isolates (Clark et al., 2012; Comas et al., 2009; Xiao et al., 2011), it is clear that there is a need for better typing methods. By leveraging the ever expanding volumes of WGS data available for many bacterial organisms, it may be possible to design novel typing methods that fulfill the criteria set by Van Belkum et al. (2007).

1.10 Overview of the thesis

Despite the decreasing costs of sequencing, WGS of bacterial isolates for epidemiological surveillance may be a long way from being economically viable option for the characterization of isolates. However, there are challenges and limitations with many conventional typing methods in terms of cost, throughput, informativeness and discrimination (Clark et al., 2012; Comas et al., 2009; Van Belkum et al., 2007; Xiao et al., 2011). Historically, typing methods have been used as a proxy for genomic data. Typing data has been used to estimate genetic similarity between bacterial isolates. However, as more sequence data are made available, it is becoming increasingly evident that typing methods can grossly underestimate or overestimate the genomic similarity between isolates (Taboada et al., 2008). Although increased sequencing efforts have expanded our understanding of the genomic variability within many bacterial species, there has been limited

application of this knowledge towards assessing current molecular typing methods and developing novel molecular typing methods.

Comparative genomic fingerprinting (CGF) is a method of comparative genomics-based bacterial characterization based on the concept that differential carriage of accessory genes can be used to generate unique genomic fingerprints for molecular typing purposes (Laing et al., 2008; Taboada et al., 2012). Our research group has successfully deployed PCR-based CGF assays for *E. coli* and *C. jejuni* (Laing et al., 2008; Taboada et al., 2012). The *E. coli* O157:H7 23-gene CGF assay (CGF23) was found to have discriminatory power comparable to PFGE and greater specificity to *E. coli* O157:H7 phage types than PFGE (Laing et al., 2008). The *C. jejuni* 40-gene CGF assay (CGF40) was shown to have high concordance with MLST and greater discriminatory power than MLST for a dataset of 412 isolates from animal, environmental and clinical sources (Taboada et al., 2012). Although the current generation of CGF assays were developed based on comparative genomic analysis of M-CGH data (Laing et al., 2008; Taboada et al., 2012), it will be possible to develop the next generation of CGF assays based on comparative genomic analysis of WGS data due to advances in sequencing technologies.

1.10.1 Goals of the thesis

We propose the development of a comparative genomic framework for the design and assessment of genomics-based conventional typing methods. Comparative genomic analysis of WGS data can be used to determine the best estimate of the population structure of a species as well as potential targets for a conventional typing method.

In order to facilitate the assessment of novel typing methods, we proposed an *in silico* typing method assessment framework. Rather than comparing typing methods against each other, typing methods would be assessed based on concordance with the underlying WGS data. Since WGS data provides the most accurate and highest resolution characterization of bacterial strains, WGS is the true "gold-standard" molecular typing method. In order to facilitate the comparison of typing methods to WGS data, we proposed the *in silico* generation of typing data from WGS data since WGS data encode all the information necessary to infer molecular typing profiles. Hence, the concordance of *in silico*-derived typing results to whole-genome phylogenies could be used as a measure of assessing the performance of a current typing method or in the development

of a novel typing scheme. Therefore, using *in silico* typing analysis, markers could be selected for inclusion in a novel typing scheme that optimize concordance with a whole-genome phylogeny as well as provide a high degree of discriminatory power.

We chose *L. monocytogenes* as the model organism for development of a novel genomics-based typing method. *L. monocytogenes* is a Gram-positive rod-shaped bacterial pathogen found ubiquitously in nature (Farber and Peterkin, 1991). Ingestion of food contaminated with *L. monocytogenes* is a major cause of human listeriosis accounting for 99% of human cases (Mead et al., 1999; Swaminathan and Gerner-Smidt, 2007). In 2008, Canada saw its worst outbreak of listeriosis resulting in 22 deaths and at least 57 illnesses (Gilmour et al., 2010; Government of Canada, 2008, 2009). PFGE was the main typing method used during this outbreak. Since the Canadian outbreak, a deadly outbreak in the USA associated with cantaloupe lead to 146 invasive illnesses, 30 deaths and one miscarriage (Laksanalamai et al., 2012) underscoring the fact that *L. monocytogenes* is an ongoing challenge in food safety. A more rapid and high-throughput method for *L. monocytogenes* subtyping could lead to more rapid characterization of and response to listeriosis outbreaks.

Although PFGE is currently the "gold-standard" method for typing *L. monocytogenes* isolates (Conly and Johnston, 2008; Graves and Swaminathan, 2001; Swaminathan and Gerner-Smidt, 2007), it is time-consuming and labour-intensive requiring standardized protocols for inter-laboratory data-sharing (Goering, 2010). Conversely, PCR-based CGF produces easily transportable typing data and is technically simple and high-throughput especially with PCR reaction multiplexing and automated gel electrophoresis instrumentation (Taboada et al., 2012). Since a CGF assay has yet to be developed for *L. monocytogenes* and due to our research group's experience in CGF assay design, we chose to develop a comparative genomic fingerprinting (CGF) assay for *L. monocytogenes* and assess its performance using our proposed *in silico* typing method assessment framework. In addition, we proposed the development of an automated pipeline for the design and *in silico* assessment of optimized multiplex PCR CGF assays from WGS data.

Chapter 2

Material and Methods

2.1 Retrieval of *L. monocytogenes* WGS data

Whole-genome assemblies for 60 *L. monocytogenes* strains from lineages I, II and III were used in the analyses described in this thesis.

2.1.1 Retrieval of Canadian outbreak-related *L. monocytogenes* WGS data

Thirty-seven Canadian outbreak-related strains were included in the design of a comparative genomic fingerprinting (CGF) assay (Table 1). WGS data for these strains was obtained from collaborators at the Public Health Agency of Canada (PHAC) National Microbiology Lab (NML) in Winnipeg, Manitoba.

Annotations for the 37 Canadian outbreak-related *L. monocytogenes* genomes were generated using RAST (Aziz et al., 2008). The GenBank format output files from RAST were parsed using BioPerl (Stajich et al., 2002) into FastA format files.

2.1.2 Retrieval of publicly available *L. monocytogenes* WGS data

Using Linux bash scripts and Perl scripts, all publicly available *L. monocytogenes* whole genome sequencing (WGS) data were downloaded in the GenBank format from the NCBI FTP site (Table 2). The GenBank files for the 23 *L. monocytogenes* strains were parsed using BioPerl (Stajich et al., 2002) into FastA format files.

2.1.3 Annotation of unannotated *L. monocytogenes* genomes

For genomes for which there were no annotations available, the genomes were annotated using RAST (Aziz et al., 2008).

2.2 Core genome phylogenetic analysis of 60 *L. monocytogenes* strains

All ORFs from *L. monocytogenes* EGD-e were BLAST searched against all 59 other *L. monocytogenes* genomes using *blastn* (Altschul et al., 1990; Camacho et al., 2009). The core genome was defined as genes present in all 60 *L. monocytogenes* genomes at a minimum 80% alignment length coverage and 80% identity (ID) at the nucleotide level. Multiple sequence alignment (MSA) of core genes was performed using MUSCLE (Edgar, 2004). The MSA of each core gene was concatenated into a concatenome of all identified core genes (Leopold et al., 2009). Gapped positions within the concatenome were removed and a phylogenetic tree was constructed using hierarchical clustering with the average linkage or uweighted pair group with arithmetic mean (UPGMA) method and saved in the Newick file format. SNPs were identified in the 60 genome concatenome and a pairwise SNP count matrix was calculated. The concatenome phylogenetic tree was generated and visualized within R (R Core Team, 2012) using the 'ape' package (Paradis et al., 2004). Low and high resolution core genome phylogeny clusters or phylogenomic clusters (PGC) were defined at 150 and 20 SNPs, respectively, within R.

2.3 Identification of homologs using pairwise protein sequence BLAST searching

Pairwise reciprocal best BLAST hit searching was performed using *blastp* (Altschul et al., 1990; Camacho et al., 2009) with the ORF amino acid sequences for all 60 *L. monocytogenes* strains to determine homologs at 85% ID and 80% alignment length coverage.

2.4 Defining a CGF marker pool from the accessory genome of *L. monocytogenes*

All ORFs within the 60 *L. monocytogenes* strains were pooled together into a CGF marker pool. Core gene ORFs and redundant accessory gene ORFs, as defined by pairwise protein sequence BLAST searching (Section 2.3), were filtered from the CGF marker pool. Redundant accessory gene ORFs were defined as ORFs already represented by an allele in the CGF marker pool. Only one allele was used to represent each accessory gene in the marker pool. Plasmid ORFs and ORFs less than 300 bp were also removed.

All alleles within the CGF marker pool were BLAST searched against the 60 *L. monocytogenes* genomes using *blastn* to determine presence rates. Alleles present in only one and 59-60 of the 60 *L. monocytogenes*, and alleles with BLAST %ID hits in 60 *L. monocytogenes* genomes between 95% and 65% were removed from the CGF marker pool. The resulting 762 alleles comprising 169 unique absence/presence patterns represented the potential CGF markers.

2.5 CGF marker set selection using the Wallace coefficient

We developed a C# application, CGF Optimizer (CGFO) (source code available at `https://bitbucket.org/peterk87/cgfoptimizer`), to find an optimal marker set solution by randomly assembling markers from unique patterns in the 762 allele marker pool given n number of tries. The 762 allele markers were collapsed down to 169 unique absence/presence patterns for

marker set generation. Each unique fingerprint represented one or more markers with the same absence/presence pattern in the 60 *L. monocytogenes* genomes.

The absence/presence patterns of each marker in a marker set were used to construct a neighbor-joining (NJ) tree for the marker set. The code for NJ tree construction was adapted from source code available for Clearcut (Sheneman et al., 2006). Marker clusters for each marker set were defined at the 95% similarity level based on the NJ tree. Using the low and high resolution PGC derived from core genome phylogenetic analysis, Wallace coefficients (W) of the marker set clusters against the low and high resolution PGC were calculated. The code for calculating the W was adapted from source code available at the ComparingPartitions website (Carriço et al., 2006). The marker sets with the highest W were saved from each iteration of random marker set generation.

A Robinson-Foulds (RF) symmetric distance (Robinson and Foulds, 1981) was calculated for each randomly-assembled marker set between the randomly-assembled marker set NJ tree and the core genome phylogenetic tree. The code for RF was based on the HashRF algorithm for rapid RF calculation (Sul et al., 2008). The Simpson's index of diversity (D) (Simpson, 1949) was calculated for each randomly-assembled marker set based on clusters defined at the 95% similarity level from the NJ tree.

2.6 Multiplex PCR assay creation

The randomly-assembled marker set with the highest Wallace coefficients with respect to the low and high resolution PGC was used for multiplex PCR assay creation. For each marker in the marker set, the largest gene with the least SNPs was used for PCR primer generation. The sequences for each marker were multiple sequence aligned (MSA) using MUSCLE (Edgar, 2004).

SNP-free primers were generated from the consensus sequence of each MSA using Primer3 (version 2.3.4) (Rozen and Skaletsky, 2000). The melting temperature, primer sequence lengths and product sizes were specified for Primer3 primer generation. Primers were created with a melting temperature between 58-62° C with an optimal melting temperature of 60° C. The acceptable length for primer sequences was 18-27 bp with an optimal primer sequence length of 22 bp.

Primers with PCR product sizes ranging from 150 to 700 bp in 50 bp increments (±25 bp size variation at each interval) were created for each marker (Figure 3.6).

All pairwise thermodynamic interactions between primers and their expected PCR products were calculated using MultiPLX (Kaplinski et al., 2005). Thermodynamic interactions were calculated based on primer to primer and primer to product maximum binding energies (ΔG). Thermodynamic interactions between primer sets were scored as low, normal or high stringency based on the default stringency thresholds used in MultiPLX. The overall thermodynamic stringency between two primer sets was set to the lowest stringency thermodynamic interaction between two primer sets.

Compatible PCR multiplexes were created based on the predicted thermodynamics using a C# application we developed, CGF Multiplexer (Figure 3.7) (source code available at `https://bitbucket.org/peterk87/cgfmultiplexer`). Grouping of primers into multiplexes was performed using exhaustive searching with pseudo-random addition of a primer set to a multiplex. Conditions for successful inclusion of a marker in a multiplex were that each marker was represented only once in the assay, PCR product sizes were staggered by either approximately 100 or 150 bp, and thermodynamics interactions between all markers within the multiplex were not below the thermodynamic stringency threshold. CGF Multiplexer would attempt to find a multiplex assay solution n number of times (user-specified) at a particular thermodynamic threshold before lowering the thermodynamic threshold for increased likelihood of finding a multiplex solution. From the 40 markers, 8 multiplexes with 5 markers each were generated using a seed of 12 for the random number generator and 3 of tries before dropping the thermodynamic stringency threshold for addition of a marker to a multiplex.

2.7 *In silico* validation of CGF assay

The 40 marker CGF assay was validated against the 60 *L. monocytogenes* strains used in CGF assay creation (Tables 1 and 2) by *in silico* PCR-based typing in Microbial In Silico Typer (MIST) (Carrillo et al., 2012; Kruczkiewicz et al., 2013, 2011). *In silico* PCR was simulated using nucleotide BLAST searching to find all potential forward and reverse primer binding sites and attempting to retrieve the likely amplicon based on those primer binding sites. BLAST searching was performed using *blastn* with default parameters and a reduced word size of 7 for

increased sensitivity. All combinations of forward and reverse primer binding sites were tested to determine potential amplicons for each marker. A positive result was reported when an amplicon was retrieved with an amplicon size within ±10% of the expected amplicon size. Otherwise, if all primer binding site combinations were exhausted and no suitable amplicon was retrieved, a negative result was reported.

Binary absence/presence fingerprints were retrieved for each of the 60 *L. monocytogenes* strains. The *L. monocytogenes* strains were clustered using complete linkage clustering of fingerprint similarity coefficients. Cluster numbers were derived at cluster definition levels of 97.5%, 95% and 90% fingerprint similarity corresponding to 1, 2 and 4 marker call differences, respectively.

In silico multiple locus sequence typing (MLST) (Salcedo et al., 2003) and multiple virulence locus sequence typing (MVLST) (Zhang et al., 2004) sequence types (ST) were assigned to each of the 60 *L. monocytogenes* strains used in CGF assay creation using MIST. All alleles for each sequence typing locus were nucleotide BLAST searched against each strain. The top allele BLAST hit for each sequence typing locus was returned. The allele numbers corresponding to the top BLAST hits for each of the sequence typing loci were used to derive the numerical sequence typing ST designation.

Using the online tool ComparingPartitions (Carriço et al., 2006), pairwise Wallace partition congruence coefficients were calculated using cluster numbers derived from core genome SNP thresholds at 20 and 150 SNPs, *in silico* CGF fingerprints at multiple cluster definition levels, *in silico* MLST ST, and *in silico* MVLST ST for the 60 *L. monocytogenes* strains used in CGF assay creation.

Chapter 3

Results and Discussion

3.1 Comparative genomic analysis of *L. monocytogenes*

The pan-genome of a bacterial species comprises the entire set, or an approximation, of the genes within the species. This includes the genes that are core to all members of the species and the accessory genes that are common to some but not all members. With the whole genome sequences of multiple members or strains from a bacterial species, it is possible to perform comparative genomic analyses comparing the genomes of these strains to look for commonalities and differences and define the pan-genome of the species. One of the principal areas of interest in comparative genomics of bacterial genomes is the determination of genetic variation such as insertions or deletions of DNA sequence (indels) and SNPs within both genes and intergenic regions that may be used in the identification of epidemiologically or clinically relevant lineages or sub-lineages in the population.

3.1.1 Defining the core genome of *L. monocytogenes*

A common comparative genomic analysis is to determine the phylogeny of strains based on the sequences of one or more core genes. The theoretical evolutionary distance between strains can be calculated based on these sequences using a nucleotide substitution model. One approach to creating a phylogeny based on more than one gene is to concatenate the aligned sequences into a

concatenome. Typically when constructing a phylogeny from a concatenome, one uses as many core gene sequences as possible to minimize the amount of bias any one gene introduces into the phylogeny while maximizing the bootstrap support of each node within the phylogeny. For the determination of the population structure of the 60 *L. monocytogenes* strains, we derived a phylogeny from the concatenome of the genes we identified as core to those 60 strains.

Homologous sequences commonly arise due to either inheritance from a common ancestor or through a gene duplication event. Homologous sequences acquired through vertical inheritance from a common ancestor are orthologous while those due to a gene duplication event are paralogous. Gene duplication events occur in the same organism rather than in an ancestor and lead to introduction of another copy of the gene into the genome of the organism. Since this duplicate gene is redundant, it may mutate and diverge in function from the original gene. In the context of determining the core and accessory genes of a bacterial species, the orthologous genes or orthologs are determined since those represent the same gene within different strains that have been vertically inherited from a common ancestor. The number of core and accessory genes estimated within a bacterial pan-genome depends on the nucleotide or amino acid identity thresholds used as well as the method used for determining orthologs. Determination of orthologous groups of core and accessory genes can be done by a variety of methods such as looking for the reciprocal best BLAST hit for a gene or by using more advanced approaches available in software like OrthMCL (Li et al., 2003).

In order to define the genes that were core to all 60 *L. monocytogenes* strains, genes that were present within all strains were identified through reciprocal best nucleotide BLAST hit searching of all *L. monocytogenes* EGD-e genes. The *L. monocytogenes* strain, EGD-e, was used as the reference strain for core genome determination since its genome was completed and manually annotated (Glaser et al., 2001). Core genes were defined as genes with alignment lengths $\geq 80\%$ and percent identities $\geq 80\%$. In total, 2314 genes were identified that met these criteria of sequence similarity and conservation. These genes were defined as the core genome of *L. monocytogenes* for the purposes of the analyses conducted in this thesis.

3.1.2 Determination of core genome phylogeny and SNP analysis

In order to determine the population structure of the 60 *L. monocytogenes* strains, the core genome phylogeny for these strains (Figure 3.1) were constructed from a concatenome of 2314 core genes of the core genome defined in Section 3.1.1. The length of the concatenome was 2,142,837 bp with gaps and 2,063,060 bp without gaps. Gap positions were removed prior to construction of the neighbor-joining phylogenetic tree as these gap positions may be phylogenetically uninformative since these positions may represent areas of low sequencing coverage.

Pairwise SNP counts between strains were determined based on the non-gapped nucleotide sites in the 2314 core genes. A core genome phylogenetic tree was obtained by hierarchical clustering of the pairwise SNP count matrix (Figure 3.1). Phylogenomic clusters were defined through "cutting" of this tree at the 20 SNP and 150 SNP cutoff levels representing high and low resolution reference clusters for CGF assay creation (Figure 3.2). There were 19 and 27 clusters generated at the 150 SNP and 20 SNP cutoff levels, respectively. The 20 and 150 SNP thresholds corresponded to a little over 1 and 2 $\log_{10}$ SNPs, respectively, in the pairwise SNP matrix (Figure 3.1).Although, for the purposes of assay design and assessment in this thesis, phylogenomic clusters were defined at 20 and 150 SNPs, phylogenomic clusters could be defined at a variety of SNP thresholds as shown in Figure 3.2.

The PGC defined at 20 and 150 SNPs were concordant with serotype, lineage and epidemiological information. A large phylogenomic cluster (blue phylogenomic cluster at 20 and 150 SNPs in Figure 3.2) of 19 and 18 lineage II serotype 1/2a strains was observed at 150 and 20 SNPs, respectively. These strains were isolated from blood, environmental, and meat samples during various listeriosis outbreaks in Canada from 1988 to 2010 (Table 1). A cluster of 3 lineage I serotype 4b strains at 20 SNPs (red phylogenomic cluster at 20 SNPs in Figure 3.2) corresponded to strains isolated during listeriosis outbreak in British Columbia in 2002 due to contaminated cheese. At 150 SNPs, these strains formed a larger cluster of 8 strains (red phylogenomic cluster at 150 SNPs in Figure 3.2) including strains isolated from a 1981 listeriosis outbreak in the Canadian Maritimes due to contaminated coleslaw (Schlech et al., 1983) and the lineage I serotype 4b strain F2365 from the 1985 listeriosis epidemic in California (Nightingale et al., 2007). A cluster of two lineage II serotype 1/2a strains (10-0812 and 10-0813; gray phylogenomic cluster at 20 SNPs in Figure 3.2) at 20 SNPs were isolated during a listeriosis outbreak in Manitoba

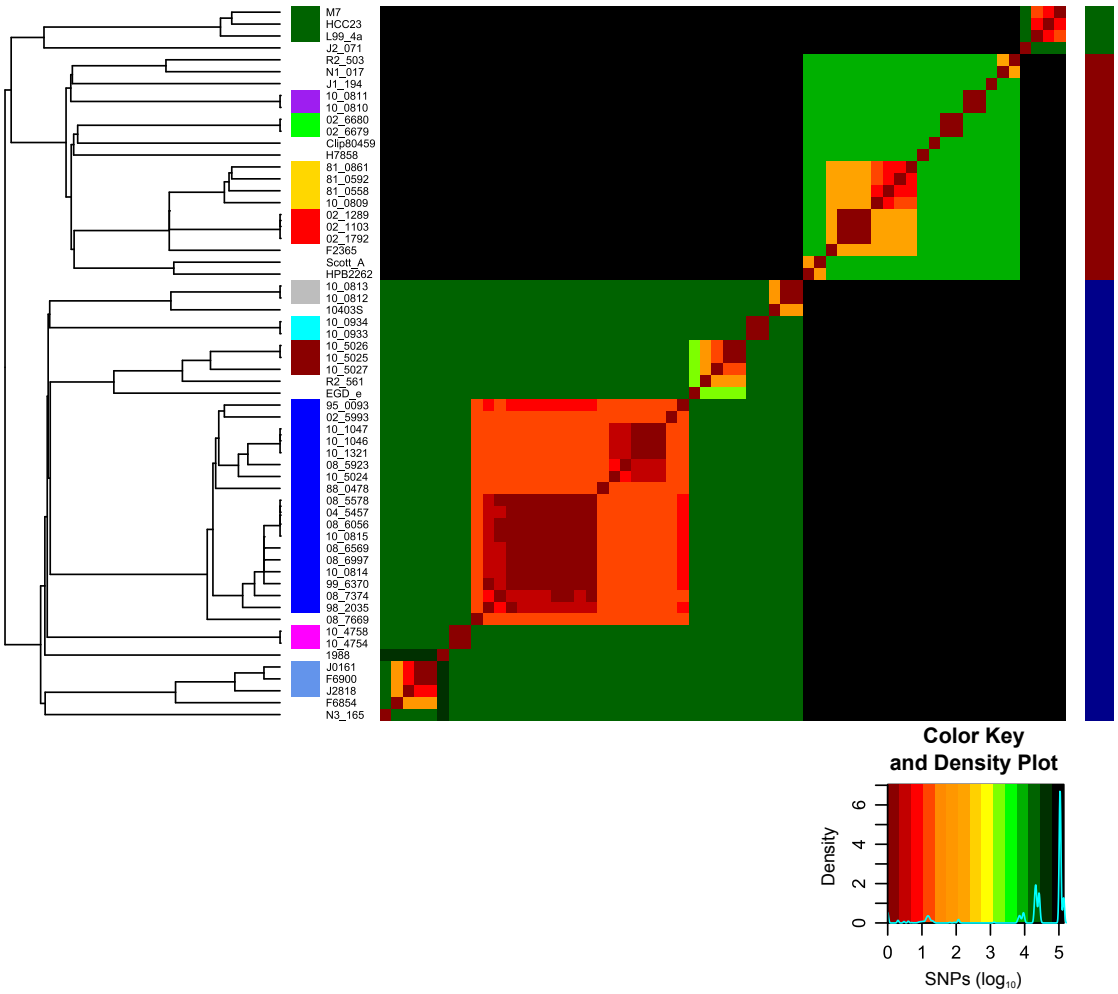

FIGURE 3.1: Log_{10} pairwise SNP counts between the 60 *L. monocytogenes* strains were calculated from the 2314 core gene concatenome. A heatmap of the log_{10} pairwise SNP counts is shown here. Hierarchical clustering of the SNP matrix was performed using the average linkage or uweighted pair group with arithmetic means (UPGMA) method. The strains cluster into three distinct lineages corresponding to lineages I, II and III of *L. monocytogenes*. Lineages I, II, and III are highlighted in red, blue, and green, respectively, to the right of the heatmap. Over 10,000 SNP differences are observed between strains from different lineages. Clusters were determined at 20 SNPs with clusters of greater than one member highlighted in a distinct colour. Cluster colours and memberships at 20 SNPs are identical to those in Figure 3.2. The relatedness of strains within these clusters can be observed in the heatmap. There are fewer than 30 SNP differences between the strains of the blue 20 SNP cluster. This is represented by a large orange and red square in the heatmap.

in 2000 due to contaminated whipping cream. Interestingly, at 150 SNPs, these two whipping cream outbreak strains formed a cluster with the lineage II serotype 1/2a strain 10403S (gray phylogenomic cluster at 150 SNPs in Figure 3.2) isolated from a skin lesion (Edman et al., 1968). These observations show that even at 150 core gene SNP differences, lineage and serotype designations may be preserved, however, epidemiological, temporal and spatial information may

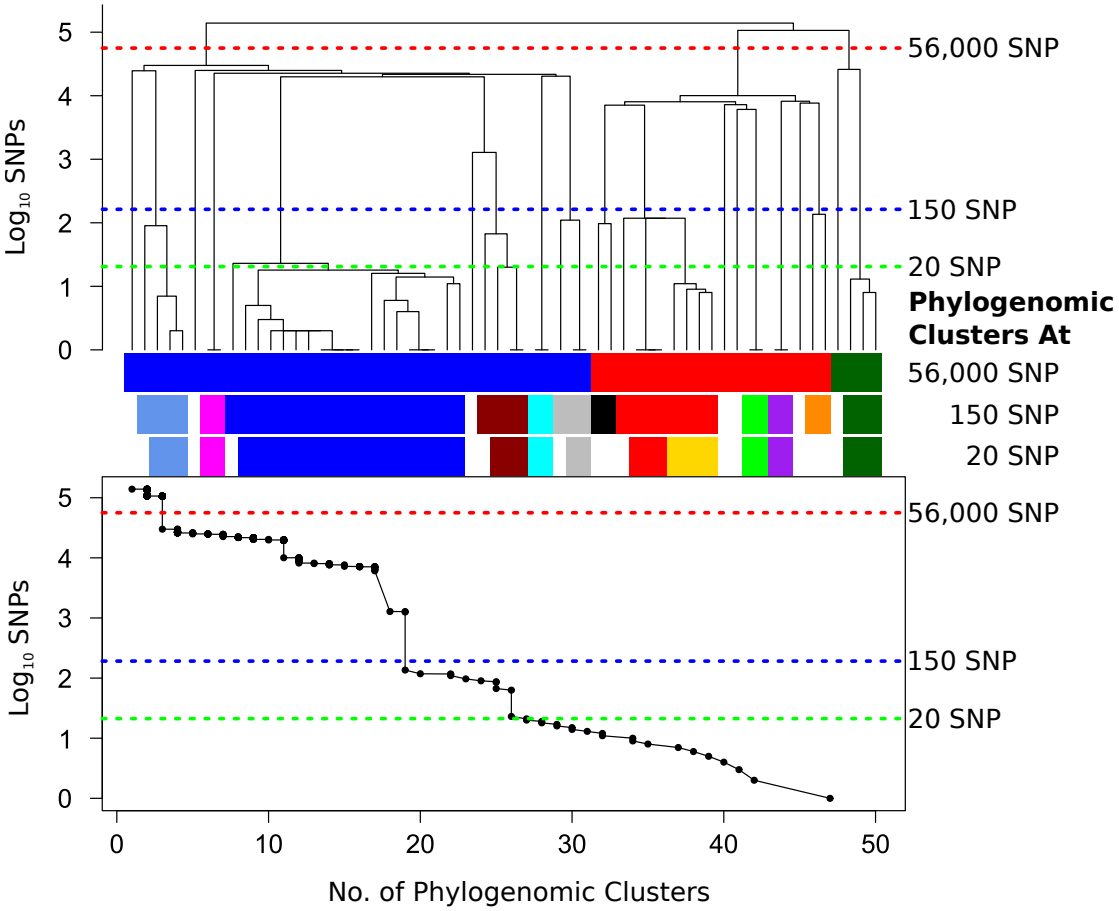

FIGURE 3.2: The core genome phylogenetic tree can be "cut" at various phylogenetic distance thresholds to produce variable numbers of phylogenomic clusters (PGC) with differing sizes and memberships. SNP thresholds at 20, 150 and 56,000 SNPs are highlighted on the core genome phylogenetic tree and the plot of number of clusters versus $\log_{10}$ SNP distance in green, blue and red, respectively. PGC defined at 20, 150 and 56,000 SNPs are shown below the phylogenetic tree. PGC with more than one member are highlighted with a colour block unique to that cluster for that SNP threshold; PGC with only a single member are not shown.

be drastically different indicating that lineage and/or serotype information may be insufficient for epidemiological investigations of *L. monocytogenes*.

The PGC defined at 20 and 150 SNPs became the reference clusters by which all typing results were compared and assessed. A typing method with high congruence to the PGC was assumed to have a high level of concordance with the underlying population structure of *L. monocytogenes*.

3.2 Optimized CGF assay design

When determining suitable accessory gene markers for a subtyping assay like CGF, one should be more stringent than when trying to estimate the number of accessory genes within the pan-genome of a bacterial species. Genes that exhibit a high level of sequence variability may be undesirable markers for a CGF assay due to the potential for SNPs within primer binding sites, which might lead to low levels of PCR product amplification and ambiguous results in the lab. Hence, it is prudent to remove any genes that may present challenges at a later stage in assay development and focus on the genes that would give unambiguous results in the lab.

3.2.1 Determination of accessory genes suitable for CGF assay design

Since a CGF assay targets genes with variable presence within a population of bacterial strains, it was necessary to determine the accessory genes of the 60 *L. monocytogenes* strains. Accessory genes were defined as genes present in at least two strains and absent in at least two strains to exclude unique, core and potentially core genes. Unique genes were defined as those present in only one strain, and were excluded since they have low information content where only a single strain would be differentiated by a unique gene. Additionally, it would not possible to design SNP-free primers with a unique gene since SNP sites cannot be determined from a single sequence. Core genes were defined as those present in all strains, and were excluded because, like unique genes, they also have low information content since no strains would be differentiated by these genes. Potentially core genes were defined as those absent in only one strain, and were excluded due to the possibility that these genes were not found due to sequencing or assembly errors.

Annotations were obtained for each strain from NCBI or, if an annotation was unavailable, through automated annotation using RAST (Aziz et al., 2008). It was necessary to annotate all Canadian outbreak-related strain WGS data (Table 1) using RAST as only the whole genome assemblies for each of these strains were available. Annotations were acquired to determine open-reading frames (ORFs) for each genome that were likely gene coding DNA sequences (CDS). All ORFs from all 60 *L. monocytogenes* strains were grouped together into a pool of potential CGF markers. Core gene ORFs were filtered out based on the core genome phylogeny analysis (Section 3.1.1). ORFs representing the same gene within multiple strains were also filtered out;

a single representative ORF for each accessory gene was left in the marker pool. At this point, 6676 ORFs were left in the marker pool. ORF presence was determined by nucleotide BLAST searching. Accessory gene ORFs with nucleotide BLAST percent identities between 65% and 95% in the other *L. monocytogenes* strains were removed reducing the number of ORFs in the marker pool to 3998. These accessory genes represented genes with ambiguous absence/presence or significant sequence variability and were removed. Only ORFs with length ≥300 bp were kept in the marker pool to improve the likelihood of finding PCR primers generating PCR products of sufficient size for the final staggered product size multiplex PCR assay. After these filtering steps, 762 ORFs remained in the marker pool from a starting total of 176,819 ORFs in all 60 *L. monocytogenes* genomes.

The remaining 762 ORFs within the marker pool represented 169 unique patterns of absence and presence across the 60 *L. monocytogenes* strains with each unique absence/presence pattern providing slightly different information on the overall relationship of the strains (Figure 3.3). Clustering of the strains based on fingerprints derived from these 169 unique patterns was congruent with the clustering based on core genome analysis. Slight gene absence/presence differences could be observed within highly homogeneous PGC suggesting potential for subtyping resolution from a CGF assay.

3.2.2 Optimized marker selection

Before designing a molecular typing assay, it is necessary to decide on whether the assay will be used for identification of particular strains or lineages or for randomly fingerprinting any strain. There are advantages and disadvantages to both approaches of typing in terms of typeability, specificity and resolution. Although an identification-based typing assay may be more suited for a well-characterized population where the distinguishing traits within that population have been thoroughly defined, this approach may fail to characterize new strains that do not possess any of the traits that the assay targets. Although the markers used in fingerprinting may not have the same biological relevance as those in an identification-based assay, a well-designed fingerprinting assay may be suited for identification purposes since a particular fingerprint may be specific to a particular subtype in the same way that a lineage-specific marker would be.

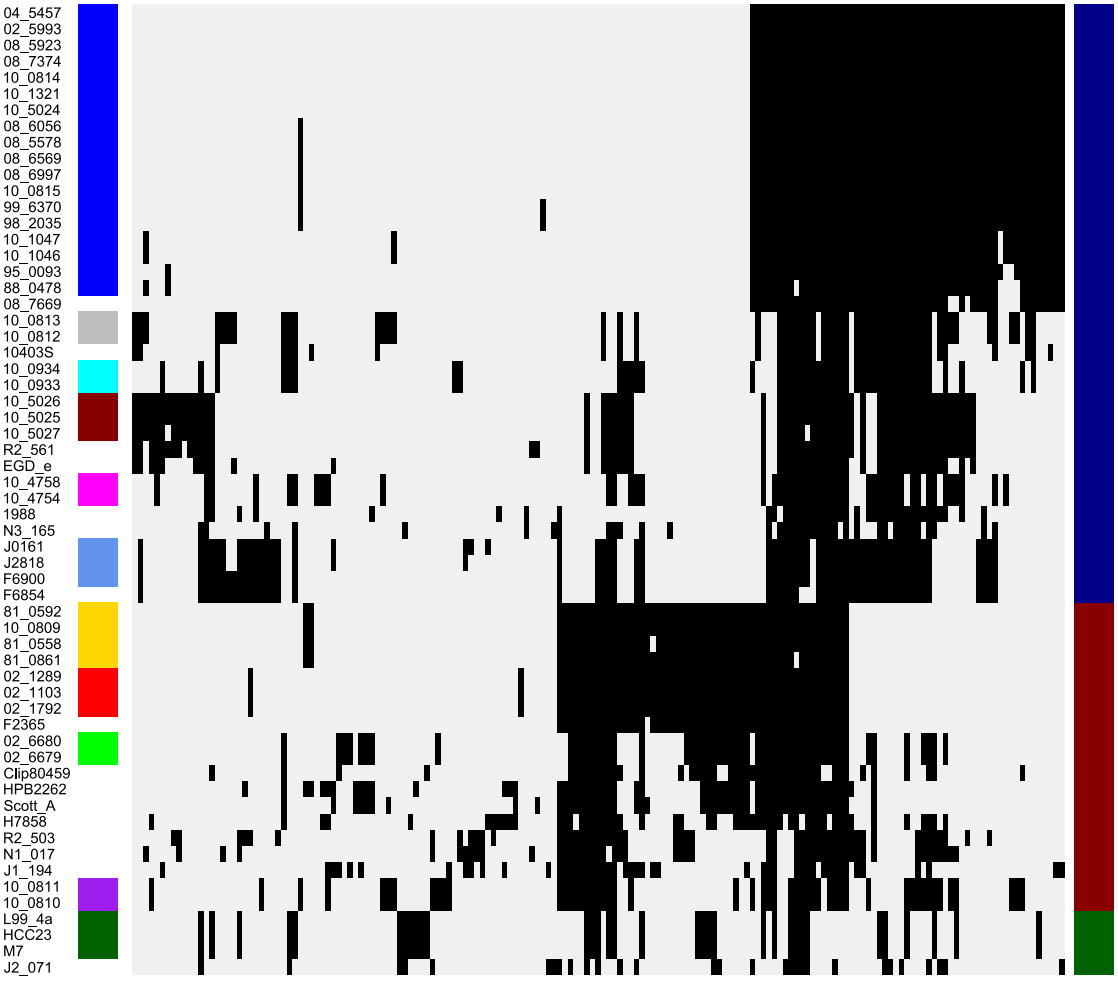

FIGURE 3.3: The 169 unique absence/presence patterns of the 762 ORFs are shown here. Core genome clusters defined at 20 SNPs are shown on the left. Lineages I, II, and III are highlighted in red, blue, and green, respectively, to the right of the heatmap. Black represents presence while white represents absence of a particular marker in a strain.

If we look at a fingerprinting-based assay, such as PFGE, we can see that nearly all strains of *L. monocytogenes* can be typed by PFGE. Every strain has a particular PFGE banding pattern and this banding pattern is its PFGE fingerprint. In some cases the fingerprint may be diagnostic for a particular lineage or clone; in other cases, the PFGE fingerprints of certain strains may lead to groupings that are too broad for epidemiological purposes. For some isolates, it may not be possible to perform PFGE due to DNA degradation, hence, these isolates may be untypeable by PFGE (Silbert et al., 2003). Greater subtyping resolution may be desired to appropriately distinguish sub-lineages for epidemiological investigations.

If one was to design a molecular subtyping assay for fingerprinting of a bacterial species, one would likely target genetic variation such as variably absent or present accessory genes or repetitive

genetic elements such as tandem repeats. Typically one would only target a small subset of markers from a pool of potential markers due to laboratory constraints such as workload, time and cost. Therefore, the selected markers should maximize the informativeness of the assay. However, if prior knowledge about the distribution of markers within a population is limited, then the selection criteria for determining which markers to target will be limited. One may simply end up picking markers at random since there may be no way to tell which markers are more informative than others. However, if one has access to a comprehensive collection of whole genome sequences for the species, then this knowledge can be used to aid the selection process so that markers can be selected that best represent the current best estimate for the population structure of the species.

With a relatively comprehensive collection of WGS data for *L. monocytogenes*, we attempted the design and creation of an optimized CGF assay for *L. monocytogenes*. Using the current best estimate of the population structure of these strains, we selected markers on the basis of how well the markers, in the context of a CGF assay, were able to infer the population structure. Some markers in combination may provide corroborating information for clustering certain strains together, while other markers may provide information for subdividing a homogeneous cluster of strains. Therefore, we generated candidate CGF assays with markers selected at random from the pre-defined marker pool in Section 3.2.1, assessed how well each candidate CGF assay was able to infer the population structure of *L. monocytogenes*, and used the best candidate CGF assay for creation of a multiplex PCR-based CGF assay.

We calculated the Wallace coefficient (W) to assess how well a candidate CGF assay was able to infer the population structure of *L. monocytogenes*. The W provides directional information between two typing methods (i.e., $W_{A\rightarrow B}$ and $W_{B\rightarrow A}$) (Wallace, 1983). This means that if two strains are in the same cluster by one method, the W can tell us the probability that these two strains will be in the same cluster using another method. In the case of assessing candidate CGF assays, the W was calculated for candidate CGF assay clusters to the PGC defined at 150 and 20 SNPs ($W_{\mathrm{CGF}\rightarrow\mathrm{phylogenomic}}$) (Section 3.1.2). The PGC were defined to enable comparison of the typing method data to WGS data. The W could range in value from 0.000 to a maximum of 1.000 representing complete disagreement and complete agreement, respectively, in one direction (e.g., method A to method B). In the context of comparing CGF assays to the PGC, a W of 1.000 would signify that the CGF assay provides identical or finer clustering of the strains than the

PGC. Conversely, a low W would mean that there are discrepancies in the clusters generated by the CGF assay compared to the PGC. These discrepancies could be due to strains being grouped into the wrong clusters or certain PGC collapsing into larger clusters. Hence, in order to find a good set of markers for the *L. monocytogenes* CGF assay, we tried to find the candidate CGF assay that maximized the W compared to the PGC.

Prior to using the Wallace coefficient (W), we evaluated the Robinson-Foulds symmetric difference (SymD) (Robinson and Foulds, 1981) as a metric for assessment of candidate CGF assays. The SymD is a tree distance measure where the number of internal node rearrangements or topological differences between two trees of the same taxa is quantified. Hence, the lower the SymD, the more topologically similar two trees are. This can be a useful measure for objectively quantifying the similarity between two phylogenetic trees. Phylogenetic trees project the population structure in two dimensions by presenting both cluster membership in addition to a second dimension represented by cluster to cluster relationships (i.e., the internal nodes of the tree). Although the results from typing methods are generally visualized in the form of a tree, there is typically insufficient information to properly assess the internal nodes of the tree. Therefore, typing methods only reliably infer cluster membership, which is a one-dimensional representation of the population structure. Because the SymD measures topological differences between the internal nodes of the respective trees, it is not an appropriate measure for assessing the concordance of a typing method to a whole-genome phylogeny. On average, a greater number of CGF markers results in fewer topological differences and a lower SymD simply due to a greater amount of information available to correctly infer the internal structure of the tree. Because we were interested in designing a CGF assay that inferred identical or a finer level of clustering than the PGC, the W was found to be a more appropriate measure.

Since the development of the framework for CGF assay design, in which we used the W to determine concordance of each candidate assay to the PGC, Severiano et al. (2011b) have gone on to publish the Adjusted Wallace coefficient (AW). The AW directly takes into account the expected W if the classifications were independent (W_i) or, in other words, occurred due to chance (Severiano et al., 2011b). Pinto et al. (2008) had previously defined the expected W under independence for method A versus B ($W_{i(A \to B)}$) to be equal to 1 minus the Simpson's index of diversity (D) of method B (i.e., $W_{i(A \to B)} = 1 - D_B$). Hence, the AW corrects for chance

agreement by following the form (Hubert and Arabie, 1985):

$$\frac{\text{Index} - \text{Expected Index}}{\text{Maximum Index} - \text{Expected Index}} \tag{3.1}$$

where the Expected Index is the expected W under independence between two classifications and the Maximum Index is assuming a maximum W of 1 (Severiano et al., 2011b):

$$AW_{A \to B} = \frac{W_{A \to B} - W_{i(A \to B)}}{1 - W_{i(A \to B)}} \tag{3.2}$$

We have not yet determined if there is a benefit to using the AW rather than the W for CGF assay design. It may be sufficient to simply rank candidate CGF assays by W since it is highly unlikely that a candidate CGF assay with a high W would be due to chance occurrence. Hence, given that the D of the PGC were sufficiently high (i.e., the PGC exhibit a high level of diversity), a high W would be indicative of a high AW; therefore, the W can be used as a proxy for the AW in this case.

Due to the combinatorial nature of selecting markers for inclusion in an assay, the number of combinations possible for generating an assay of k markers from n potential markers increases factorially with increasing k and/or n. This means that for an assay consisting of 40 markers with a marker pool of 169 markers with unique patterns, there would be $\binom{169}{40}$ or 1.052×10^{39} combinations. Assuming that assessing the performance of marker set combinations against a core genome phylogeny could be done at a rate of 1,000,000 per second, exhaustive searching of all possible combinations for the optimal set of markers would take 3.17×10^{25} years. Therefore, an approximate approach was necessary. Although there may be more efficient algorithms for quickly determining the best possible set of markers given a pool of potential markers, the goal was to select an optimized set of markers rather than try to find the best set of markers, hence, for our purposes, a brute force method of searching was implemented and tested.

Marker selection and optimization was automated using a brute force approach in a C# application we developed called CGF Optimizer (CGFO) (screenshot shown in Figure 3.5; source code available at `https://bitbucket.org/peterk87/cgfoptimizer`). To assess the agreement of the subtyping results from the candidate marker set with respect to the PGC, the W was calculated between clusters generated by the candidate marker set and clusters derived from core genome analysis. Neighbor-joining (NJ) clustering was performed using a fingerprint similarity distance

matrix for each candidate marker set. Clusters for each candidate marker set were defined at the 95% fingerprint similarity level from the NJ clustering. A similarity threshold of 95% was used to ensure that there were at least 3 or more loci differences between strains belonging to different PGC. Two sets of PGC were defined at 150 and 20 SNPs in order to establish reference low and high resolution clusters, respectively. Clusters generated by the candidate CGF assay would be required to provide finer subtyping resolution than the PGC at 150 SNPs while providing a similar or finer level of subtyping resolution relative to the higher resolution PGC at 20 SNPs.

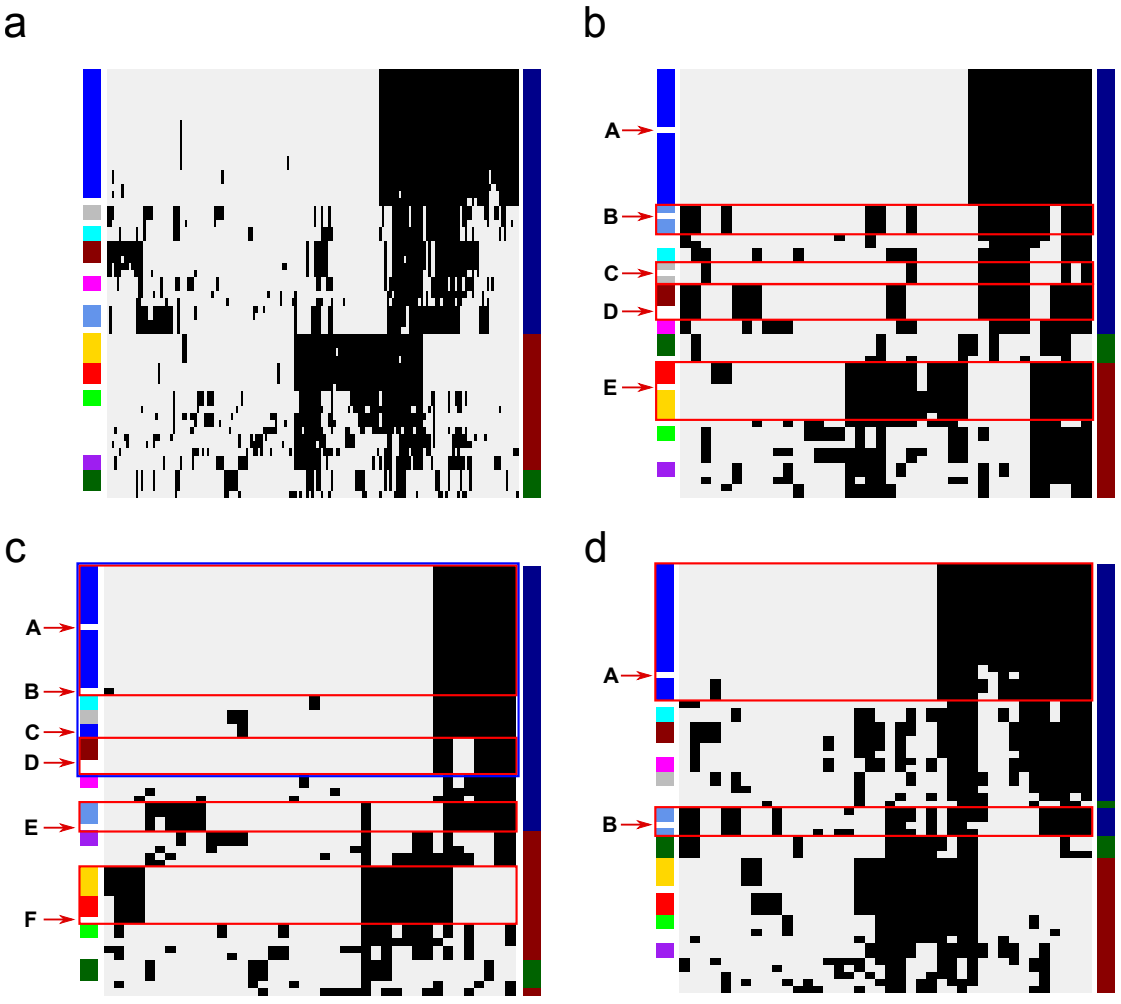

FIGURE 3.4: The absence/presence profiles of a) the entire marker pool, b) an 'average' marker set, c) a 'poor' marker set and d) the 'good' marker set are shown here. Features of interest in each figure are highlighted with arrows and/or coloured boxes. Core genome clusters defined at 20 SNPs are highlighted to the left of each absence/presence profile. Lineages I, II, and III are highlighted in red, blue, and green, respectively, to the right of each absence/presence profile. Black represents presence while white represents absence of a particular marker in a strain.

The pool of 762 accessory gene ORFs determined to be suitable CGF assay targets (Section 3.2.1) were used for CGF assay design. These 762 ORFs consisted of 169 unique absence/presence

patterns (Figure 3.3). The marker pool was reduced down to those 169 unique patterns using the most suitable ORF for CGF assay creation for each pattern. Candidate marker sets of 40 CGF markers were generated from randomly selecting markers from the pool of 169. Forty markers were selected for the *L. monocytogenes* CGF assay to maximize the information content with an assay design that could be deployable in practice in the laboratory context.

In order to highlight the importance of optimized marker selection, the differences between subtyping results of an 'average' marker set of randomly chosen markers, a 'poor' marker set and a 'good' marker set are examined with respect to PGC (Figure 3.4). Since the Wallace coefficients of two different marker sets may differ significantly, it is important to highlight the basis for this difference in the subtyping results of each respective marker set.

Choosing markers at random from the candidate patterns (Figure 3.4a), one may generate a marker set with an absence/presence profile as shown in Figure 3.4b. However, given the goals of producing an assay with resolution into highly homogeneous clusters, there are several instances where this particular set of markers fails (Figure 3.4b *A*–*E*). Strain 08-7669 has an identical fingerprint to the strains in the blue phylogenomic cluster despite the fact that this strain belongs to another phylogenomic cluster (Figure 3.4b *A*). Though this strain and the strains of the blue phylogenomic cluster are highly similar (< 100 SNPs), if subtyping resolution into this cluster is desired, this set of markers would fail to provide the necessary level of resolution to differentiate these strains. Lack of subtyping resolution is apparent in four other instances (Figure 3.4b *B*, *C*, *D* and *E*) where differences at the core genome level are not represented at the subtyping level. This marker set is an example of a marker set that could benefit from optimized marker selection since there are numerous cases of insufficient subtyping resolution between highly similar strains and clusters.

In the worst case scenario, randomly selecting markers may result in a marker set that provides very poor overall concordance to the PGC (Figure 3.4c). The Wallace coefficients for this marker set compared to the PGC defined at 150 and 20 SNPs were 0.455 and 0.357. Strain 08-7669 that belongs to its own phylogenomic cluster has an identical fingerprint to some of the strains in the blue cluster (Figure 3.4c *A*). There are several other similar instances of insufficient subtyping resolution in the subtyping results of this assay (*B*–*F*). A major deficiency with the subtyping resolution of this assay is highlighted by a blue box encompassing points of interest *A*–*D* in Figure 3.4c. The strains highlighted by the blue box would be grouped into the same cluster at a

95% fingerprint similarity threshold although these strains belong to multiple PGC. These cases of insufficient subtyping resolution demonstrate why this would be a poor marker set and how one could potentially infer incorrect relationships between strains from the results of such an assay.

Using optimized marker selection based on maximizing the Wallace coefficient of a candidate marker set to reference PGC, it is possible to find a set of markers that is highly concordant with the phylogenetic clusters. The absence/presence profile of such a marker set is shown in Figure 3.4d. The Wallace coefficients for this marker set compared to the PGC defined at 150 and 20 SNPs were 1.000 and 0.962 signifying a high level of agreement between the clusters of this marker set and the PGC. Looking at the underlying absence/presence data, it is clear that this marker set provided the required subtyping resolution into some of the highly homogeneous PGC. Strain 08-7669 had a distinct fingerprint compared to other highly similar strains within the blue core genome cluster (Figure 3.4d *A*). This is an improvement from both the 'average' and the 'poor' marker sets (Figure 3.4b and 3.4c, respectively) where this strain had a fingerprint identical to strains in the blue core genome cluster. There were also strains (10-1046, 10-1047, 88-0478 and 95-0093) within the blue core genome cluster with fingerprints that varied by one or two loci. These unique fingerprints could provide potential subtyping resolution into a cluster of highly similar strains with < 20 SNPs between them. The only discrepancy between the clusters formed from this marker set and the PGC involved strain F6854 and the light blue core genome cluster (Figure 3.4d *B*). Strain J0161 had 3-4 loci differences to the other strains, J2818 and F6900, in the light blue core genome cluster illustrating another instance of potential subtyping resolution. In the core genome phylogeny, F6854 was the outlier in the cluster of J0161, J2818 and F6900 with over 100 SNPs compared to the other strains. According to the core genome phylogeny, J0161 and F6900 should share very similar fingerprints, but F6854 was more similar to J2818 and F6900 than J0161. However, this was a minor discrepancy and an example of potential over interpretation of the relationships between strains based on CGF fingerprints since at a 95% fingerprint similarity threshold, F6854 and J0161 would form their own clusters thus provide finer level clustering than the PGC at 20 SNPs. Overall, a high level of subtyping resolution was achieved with this set of markers through optimized marker selection.

In order to generate the set of markers which would become the *L. monocytogenes* CGF assay (Figure 3.4d), we conducted a brute force search using CGFO examining a sufficiently large

number of marker sets to maximize the Wallace coefficients in comparison to the low and high resolution PGC. In total 25 million candidate marker sets were generated in order to find the marker set with Wallace coefficients of 1.000 and 0.962 to the low and high resolution clusters, respectively, at the 95% fingerprint similarity cluster definition level. As detailed earlier, the absence/presence profile of this set of markers shows a high level of agreement with the core genome phylogeny while providing a high level of subtyping resolution.

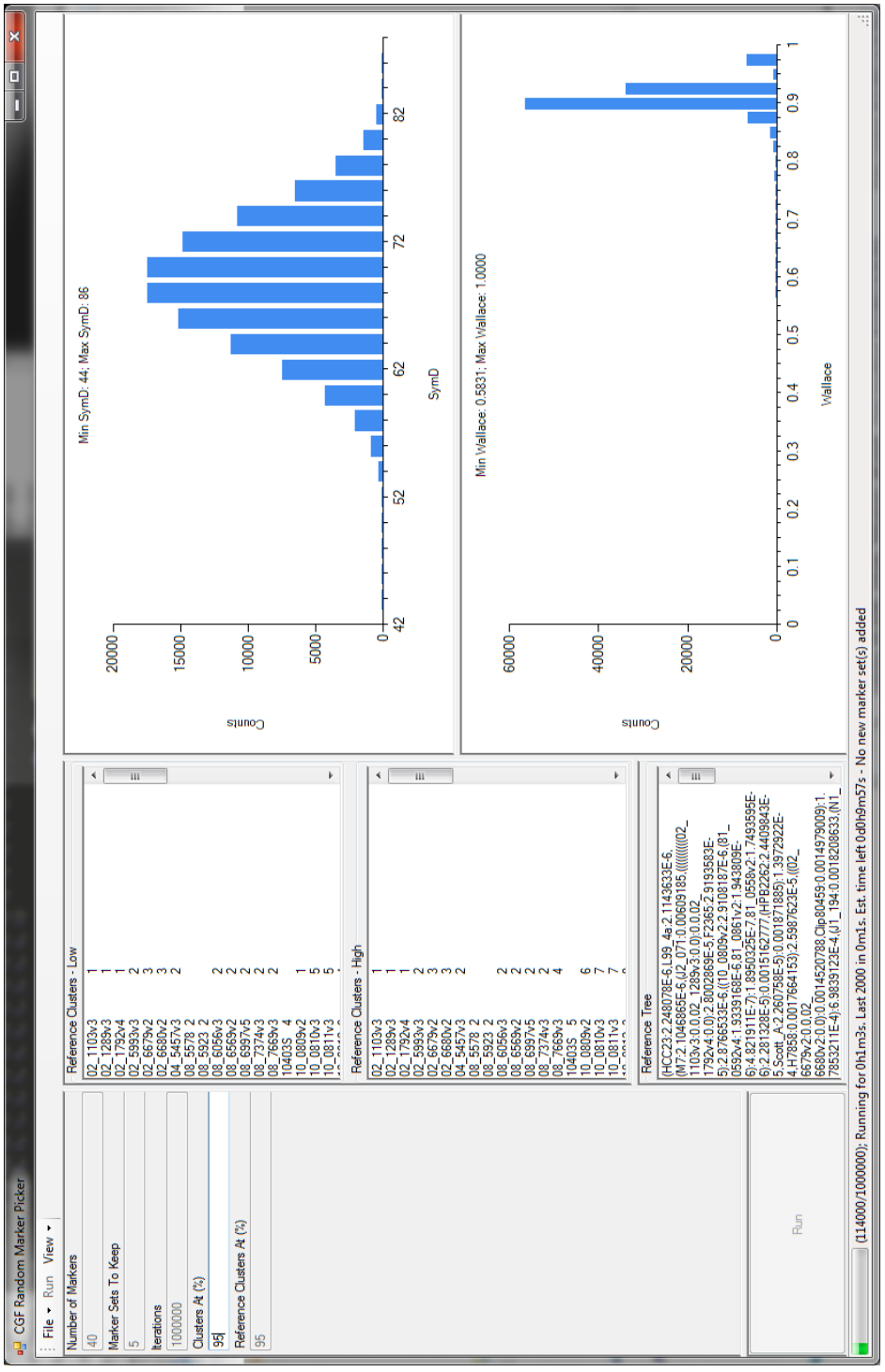

FIGURE 3.5: A screenshot of CGF Optimizer (CGFO) generating and assessing marker sets from a marker pool is shown here. The user supplies the list of markers in the marker pool and their absence/presence patterns, the number of markers to include in a marker set and the reference clusters. CGFO will randomly generate marker sets and assess the discriminatory power and concordance to the reference clusters of each of these marker sets. The marker sets with the highest discriminatory power and concordance to the reference clusters will be reported back to the user.

3.2.3 Multiplex PCR assay design

Once the optimized CGF markers were chosen it was necessary to design a multiplex PCR assay that would target those markers. Multiplexing of primers would reduce laboratory workload and cost. Less reagents are used and fewer PCRs are run. However, trying to determine how best to multiplex the primers for all of the markers in an assay in order to minimize undesired thermodynamic interactions may require extensive and costly testing in the lab. Fortunately, there are many software tools to aid in creating and grouping primers into thermodynamically-favourable multiplexes that minimize the chances of unwanted primer-primer or primer-PCR product interactions. With the proposed *L. monocytogenes* CGF assay, multiple sets of primers producing PCR products of various sizes were created for each marker if possible. The thermodynamic interactions between all primers and PCR products were determined. Primers with favourable thermodynamic interactions were grouped together into multiplexes using a program we developed called CGF Multiplexer (source code available at `https://bitbucket.org/peterk87/cgfmultiplexer`).

To facilitate the creation of primers free of SNPs, the consensus sequence of each CGF gene marker was derived from a MUSCLE (Edgar, 2004) multiple sequence alignment (MSA) of the marker sequence from each of the *L. monocytogenes* strains in which the gene was found. Primer sequences with SNPs would require degenerate primers and would complicate the thermodynamic interactions within a multiplex PCR reaction; therefore, SNPs within primer sequences were avoided by using consensus sequences for primer design. Primer3 (Rozen and Skaletsky, 2000) was used to generate the non-degenerate primers at all possible PCR product sizes within a range from 150-700 bp at intervals of 50 bp from the consensus sequence of each marker (Figure 3.6). Primers producing a range of PCR product sizes for each marker were generated to enable staggering of PCR products by 100-150 bp. This staggering of PCR products would allow for unambiguous differentiation of the presence or absence of a marker within a multiplex.

All possible thermodynamic interactions between PCR primers and products were calculated using MultiPLX (Kaplinski et al., 2005). There were five thermodynamic interactions computed in pairwise fashion between all PCR primer sets of all markers within the CGF marker set. These thermodynamic interactions were the maximum binding energy (ΔG) between the two primers including the 3' ends of both primers, the 3' end of one primer and any region of another primer,

any region of different primers, the 3' end of one primer with any region of a PCR product, and any region of a primer with any region of a PCR product (Kaplinski et al., 2005).

Since MultiPLX or any other software did not contain the necessary logic to multiplex the primers we generated for each marker into multiplexes of our choosing, we developed a C# program, CGF Multiplexer, to generate the multiplexes. CGF Multiplexer program was written to place all 40 markers into 8 multiplexes of 5 markers while ensuring the highest achievable thermodynamic compatibility of PCR primers and PCR product staggering within the multiplexes (Figure 3.7). The theoretical thermodynamic compatibility of each different thermodynamic interaction between primer sets was defined according to the default settings of MultiPLX. The lowest compatibility thermodynamic interaction of the five computed interactions was used to determine the thermodynamic compatibility between two PCR primer sets of two markers for multiplexing purposes. CGF Multiplexer would randomly try to assign marker primers to a candidate multiplex. If the primer set was thermodynamically compatible with the other primer sets already in the marker set it would move onto adding the next primer set. If all slots within the multiplex were filled, it would move onto trying to generate the next multiplex until all markers had been assigned to a multiplex. If a particular set of choices resulted in a "dead-end" and a multiplex PCR solution could not be generated, the program would retrace its steps and try to fit a different primer set into the appropriate multiplex until a multiplex PCR solution could be produced.

Using CGF Multiplexer, a multiplex PCR solution was found with high to average thermodynamic compatibility between primers sets within multiplexes (Table 5). Theoretically, this multiplex solution should minimize potential multiplex PCR issues such as primer-primer interactions. However, to determine if this multiplex solution is able to produce adequate results within the lab, further testing would be required with a panel of strains for which the CGF fingerprints are known ahead of time.

Typically, if a multiplex PCR is not generating the expected results due to primer-primer interactions, optimization of the multiplexes would be required in the lab by trying different primer combinations or reaction conditions. However, using a program like CGF Multiplexer, a new multiplex PCR targeting the same markers could be created and tested *in silico* using more stringent parameters. It may be cheaper and quicker to order more primers that show theoretical

thermodynamic compatibility *in silico* rather than trying different combinations of primers and reaction conditions in the lab.

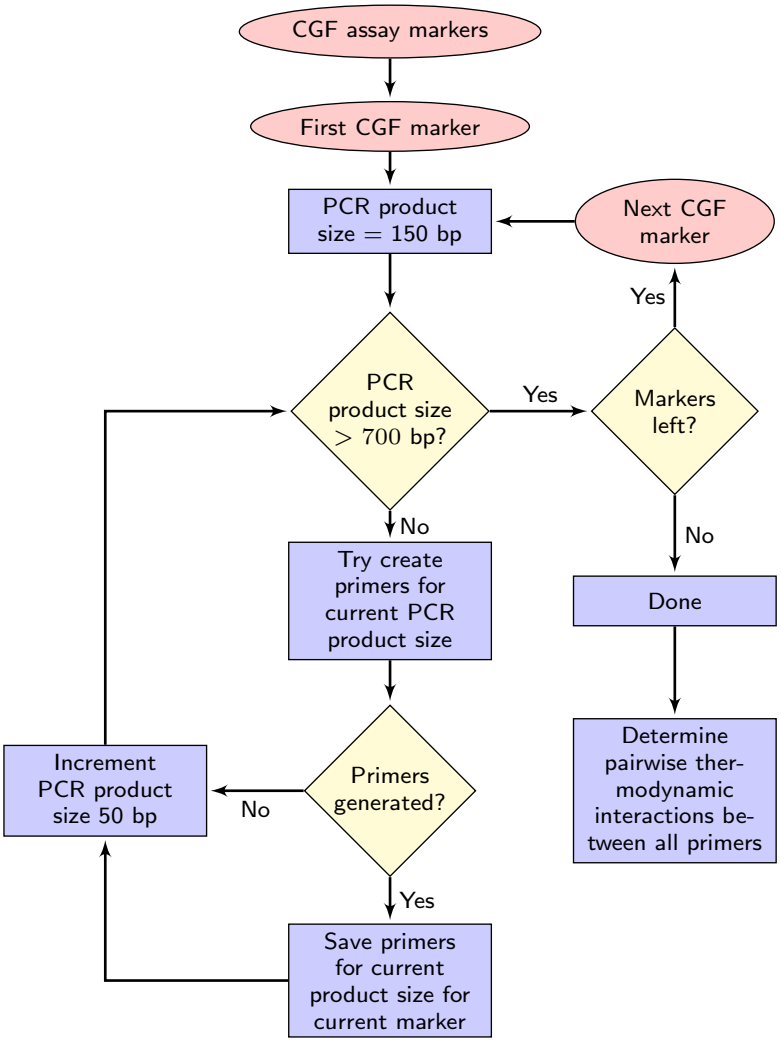

FIGURE 3.6: A flowchart for the creation of staggered product size PCR primers for each CGF marker is shown here.

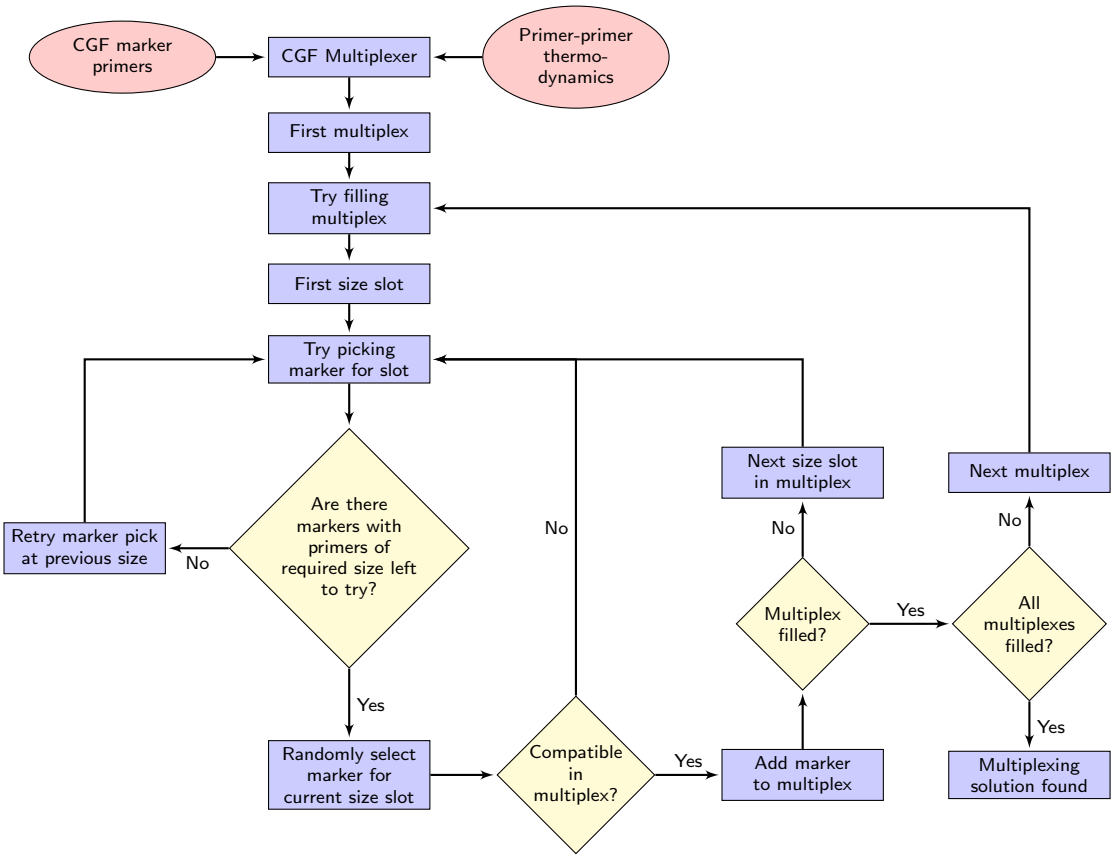

FIGURE 3.7: A flowchart showing how the CGF markers were multiplexed is shown here.

3.3 *In silico* typing using WGS data

Although comparative analyses of bacterial pathogen WGS data could expedite development of more informative molecular typing assays (Hall et al., 2010), for some priority human bacterial pathogens, such as *M. tuberculosis*, WGS may soon become the sole method for molecular typing (Schürch and van Soolingen, 2012). Therefore, as generation of WGS for many bacterial pathogens becomes more common, relating WGS data to historical typing data will be of great importance from an epidemiological standpoint.

Since the WGS data for an organism encompasses the sum total of all genetic data available for that organism, it should be possible to derive the molecular typing profiles from this WGS data using computer-based or *in silico* approaches. Generation of *in silico*-derived typing data would allow newly sequenced strains to be placed within an epidemiological and historical context through linkage with typing databases such as PubMLST (Jolley et al., 2004) or PulseNet (Swaminathan

et al., 2001). This translation of high-resolution WGS data into lower resolution conventional molecular typing data would allow new strains to be related to previously characterized strains. Therefore, *in silico* typing of WGS data could prove to be highly valuable in the context of epidemiological investigations.

In the context of an outbreak, every outbreak-related isolate may be sequenced to aid in an epidemiological investigation. However, with epidemiological surveillance, generation of WGS for every isolate may not be economically feasible. Therefore, there will always be a demand for more informative, cheaper and more rapid typing methods for surveillance of priority human pathogens. As a greater number of genome sequences become available, new molecular markers will be discovered for creation of novel next-generation typing schemes.

WGS is the true gold standard for molecular characterization of microbes offering not only the highest level of resolution for differentiating strains, but all of the information necessary to determine conventional molecular typing profiles. Consequently, WGS analysis has become increasingly prominent in the public health response to pathogens, enabling characterization of organisms at an unprecedented level of resolution. As the cost of sequencing continues to decline, the applicability of WGS will broaden to a larger scope of strains and, in the near future, will likely become the method of choice for characterization of all microbes.

In the context of public health, a significant gap currently exists in that although WGS analysis has been shown to be viable in the context of public health events such as outbreaks (Gilmour et al., 2010), it may not yet be possible to perform WGS in the context of epidemiologic surveillance. Thus, while ever-expanding, WGS datasets still comprise only a fraction of the historical data contained in public repositories of molecular typing data, such as PulseNet (Swaminathan et al., 2001) and PubMLST (Jolley et al., 2004), and this is not likely to change in the immediate future. Establishing links between WGS data and the data contained in molecular epidemiology databases will be critical as we transition from a paradigm involving molecular typing to one that is based on WGS. In particular, there is the opportunity for the utilization of WGS data as a framework for comparing the performance of existing molecular typing methods or for assessing and validating novel methods (Carrillo et al., 2012).

In order to facilitate the generation of molecular typing data from bacterial WGS data, we developed Microbial *In Silico* Typer (MIST).We have recently reported results obtained using MIST as part of a framework for assessing the performance of existing molecular typing methods

for *C. jejuni* and *C. coli* using WGS data (Carrillo et al., 2012). Additionally, we have recently presented the full implementation of MIST and demonstrated the utility of MIST analysis using finished, closed and draft assembly genome sequence data from priority human pathogens, and shown through comparison with published and experimental data that MIST produces accurate *in silico*-derived molecular subtyping data for a variety of molecular typing approaches (Kruczkiewicz et al., 2013).

MIST was designed to provide a complimentary approach to existing tools by allowing users the flexibility to analyze draft WGS data and produce *in silico* typing profiles for traditional typing schemes, assays based on modifications of these traditional schemes, and assays based on novel typing schemes. These *in silico* typing profiles can be linked to historical typing data from public databases and to other available metadata on each genome analysed, such as phenotypic, clinical, or epidemiological information, enabling the user to examine associations between the sub-typing data and their underlying metadata. Since MIST allows the user to design and test schemes based on approaches that may be impractical to perform *in vitro*, extended MLST schemes or hybridization-based methods targeting hundreds of loci can be tested that would otherwise be subject to sensitivity/specificity issues when deployed in the lab. Furthermore, the MIST platform enables comparison between typing methods through the simultaneous generation of *in silico* typing results for a variety of different methods.

3.3.1 Assumptions and limitations of *in silico* assay design and assessment

Although *in silico* PCR-based typing of bacterial organisms had high concordance with lab generated typing data (Kruczkiewicz et al., 2013), *in vitro* typing results may vary significantly from those derived by *in silico* analysis due to sequencing errors or poor sequence quality. Consequently, there may be a significant difference in the *in silico* performance of a typing method compared to the *in vitro* performance of the same method. Additionally, although PCR primers may have theoretically favourable properties or PCR primers are grouped into theoretically thermodynamically favourable multiplexes, a PCR-based assay may not generate the desired results in the lab due to complex primer-primer interactions or primer-product interactions.

For the *in silico L. monocytogenes* CGF40 assay, the clusters used for typing method design and assessment were defined at 20 and 150 SNPs within the core genome concatenome phylogeny (Figure 3.2). However, other possible thresholds representing the population structure of *L. monocytogenes* at different levels of resolution could have been used. Therefore, if a novel typing method is expected to recapitulate population structure, phylogenomic clusters should be defined at a level of resolution desired for the novel typing method (e.g., lineage differentiation or discrimination between isolates from different outbreaks).

Although a typing method could be designed with high concordance to the population structure of the target organism (e.g., CGF40 assay for *L. monocytogenes*), the framework described in this thesis could be used to design a typing method with high concordance to other forms of classification. For example, a typing method could be generated to have high concordance to epidemiological data, clinical data or phenotypic data. In order to generate a typing method with high concordance to other classifications, the desired reference classifications rather than phylogenomic clusters would need to be specified during typing method design and assessment.

Designing a typing assay targeting accessory gene absence/presence relies upon an observable level of recombination leading to insertions and/or deletions of DNA sequence (indels) within the genomes of the bacterial population of interest. Therefore, an assay targeting gene absence/presence would not be appropriate for bacterial organisms with little to no recombination resulting in indels, such as *Salmonella enterica* serovar Enteritidis (Campioni et al., 2012; Olson et al., 2007). Hence, an assay targeting SNPs or VNTR regions would likely be more appropriate for these organisms. Our framework for *in silico* design and assessment of typing assays could still be applied for determination of the set of molecular markers that provide the greatest discriminatory power and concordance with the expected clusters or classifications.

The selection of strains used in the generation of a typing method should be representative of the overall population to be targeted by a typing method. On the one hand, undersampling of certain subsets of the population could result in an assay that does not adequately classify and discriminate isolates belonging to these subpopulations. On the other hand, oversampling of certain subsets of the overall population could bias the typing assay towards the classification and discrimination of those subpopulations at the expense of others. Hence, representative and appropriate sampling of whole-genome sequenced strains for typing method design and assessment

is required for the generation of an assay that is both discriminatory and produces results that are concordant with the expected clusters or classifications.

WGS data contains all information necessary for inferring molecular typing data, which can be generated through *in silico* typing analysis of WGS data. This is especially useful when one has the WGS, but lacks molecular typing data for their strains of interest. Although the CGF assay described in this thesis was validated using *in silico* typing of *L. monocytogenes* WGS data, *in silico* validation is not a substitute for laboratory validation of a typing method. In order for the CGF assay to be considered for use in the lab, the assay would need to be validated in the lab, and compared against the current gold-standard method for typing of *L. monocytogenes*, PFGE. The *in silico* framework for molecular typing assay design and assessment described in this thesis is not aimed as a replacement for in-lab validation. Rather, this framework enables for rapid generation of assays that, having a theoretically high level of performance, would warrant full scale in-lab validation.

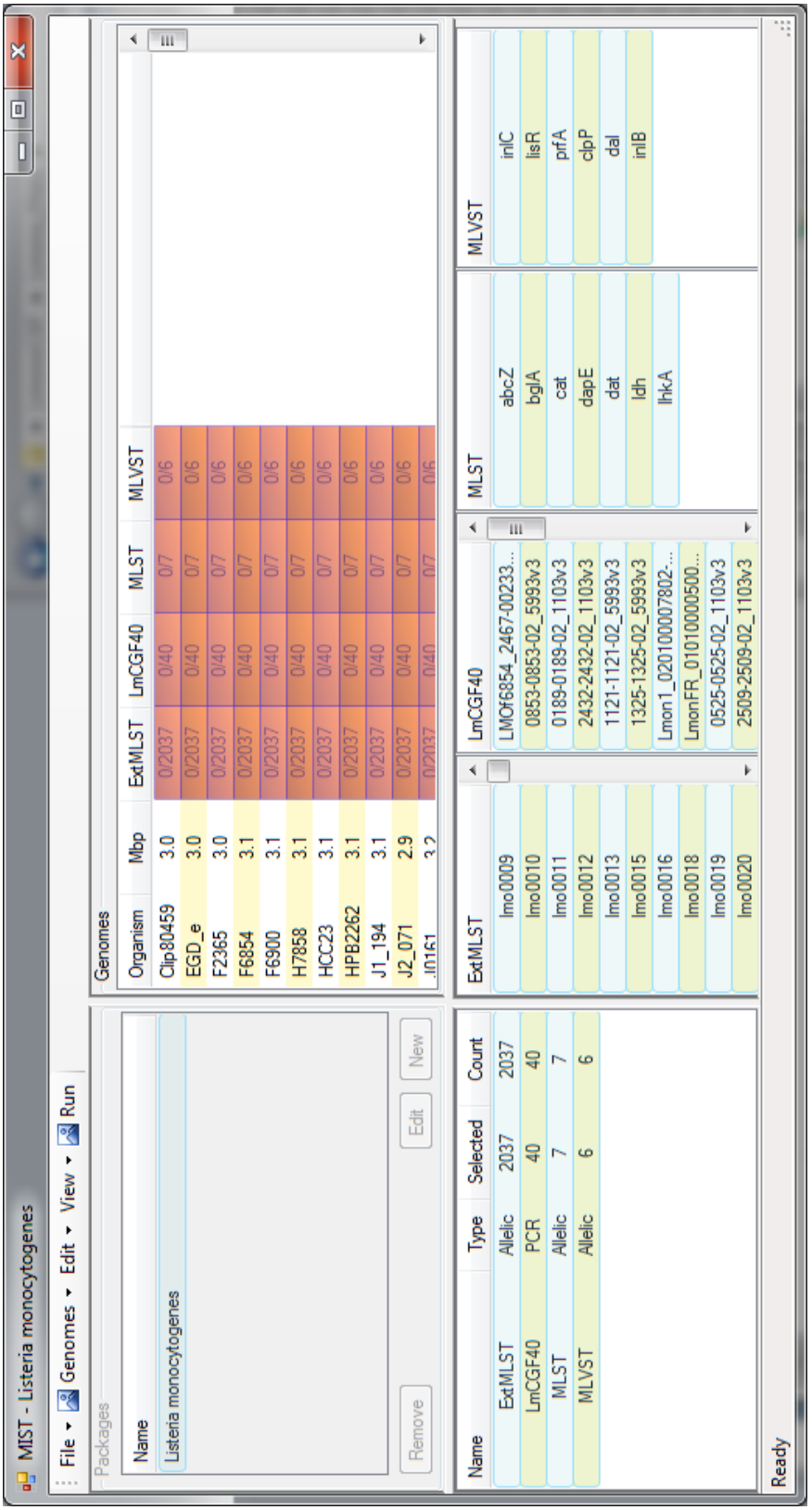

Figure 3.8: A screenshot of MIST is shown here.

3.3.2 *In silico* validation of the *L. monocytogenes* CGF40

In silico MLST, MVLST and CGF40 typing data were generated for the 60 *L. monocytogenes* strains from analysis of WGS data using the MIST program (Carrillo et al., 2012; Kruczkiewicz et al., 2013). MIST facilitates generation of *in silico* typing data for a variety of genotyping methods including PCR-based methods such as CGF and MLVA, sequence-typing-based methods such as MLST and probe-based methods such as comparative genomic hybridization (CGH). *In silico* determination of sequence typing data correlated highly with data obtained in the lab with over 98% concordance for *Campylobacter* MLST allele determination (Carrillo et al., 2012). For the *Campylobacter* CGF40, a multiplex PCR assay for subtyping *C. jejuni* and *C. coli* (Taboada et al., 2012), the concordance between *in silico* results derived from MIST analysis of WGS data and *in vitro* results from analysis of isolates in the lab was found to be 92.8% (Kruczkiewicz et al., 2013). As sequencing technologies improve in accuracy and read lengths, it will be possible to produce higher quality whole genome assemblies thereby allowing more accurate determination of *in silico*-derived subtyping information.

Clusters were defined from the *in silico* typing data in order to enable comparison of the various methods against the clusters derived from core genome analysis at 20 and 150 SNPs. Various fingerprint similarity thresholds were used to define clusters from the *in silico* CGF40 typing data. Sequence type (ST) designations were treated as cluster numbers for MLST and MVLST.

3.3.2.1 *In silico* CGF40

In silico L. monocytogenes CGF40 typing data were generated for 60 *L. monocytogenes* strains (Figure 3.9) using MIST *in silico* PCR-based typing. The *in silico* typing results for the *L. monocytogenes* CGF40 were identical to the expected absence/presence profiles given the absence/presence fingerprints of each marker defined during pan-genomic analysis (Section 2.4).

All primer binding sites for each marker were determined using nucleotide BLAST with a reduced word size of 7 for increased sensitivity. Presence of a marker was defined by whether a PCR amplicon of the expected size (with a size tolerance) could potentially be produced from the theoretical primer binding sites of the forward and reverse primers for that marker. Absence of a marker was defined as the inability to produce an amplicon with an appropriate size.

Clustering of the 60 *L. monocytogenes* strains based on the *in silico Listeria* CGF40 results was performed using UPGMA clustering of the absence-presence fingerprint similarities. Cluster numbers were defined at fingerprint similarities of 100, 97.5, 95 and 90% corresponding to 0, 1, 2, and 4 loci differences, respectively (Table 6).

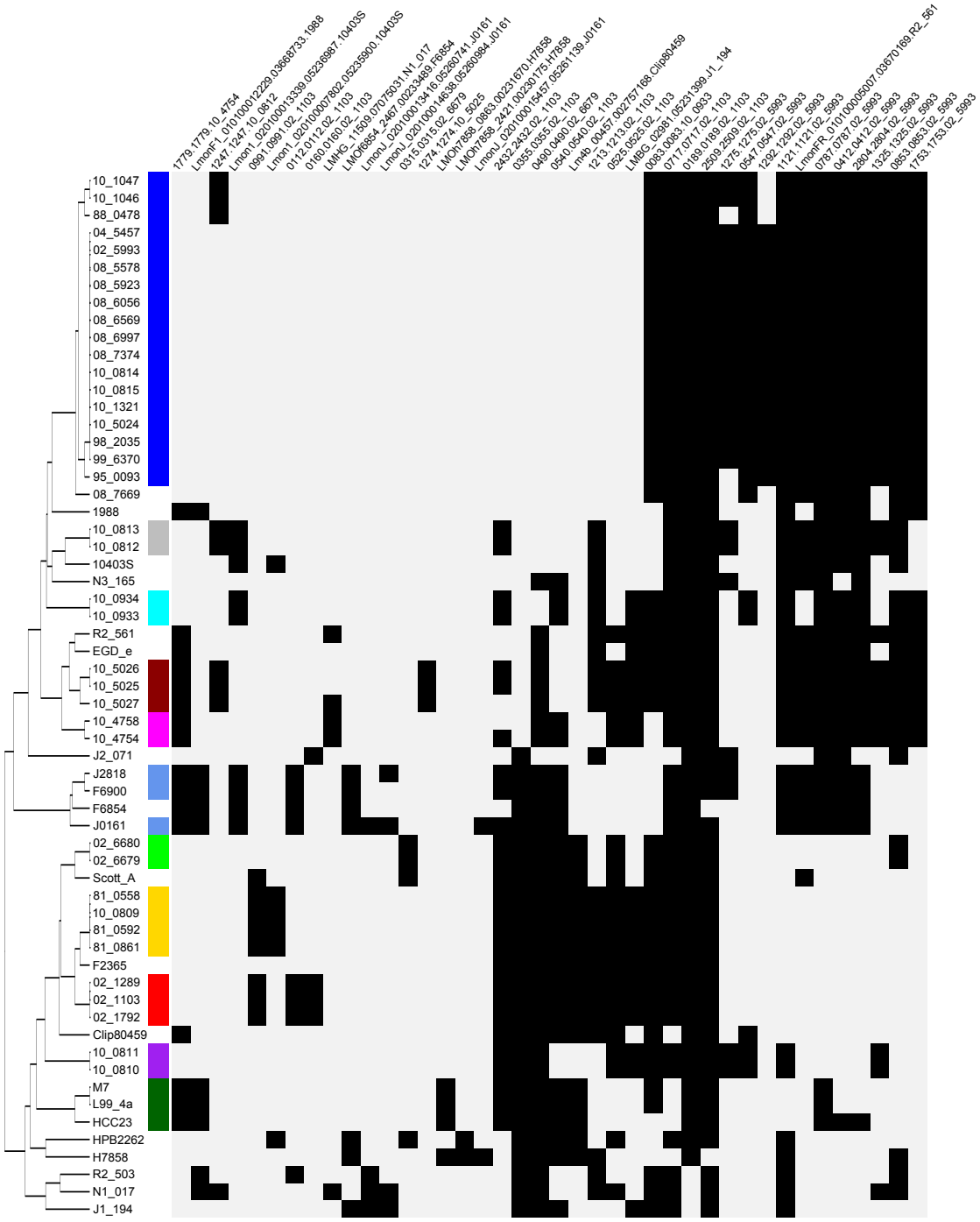

FIGURE 3.9: The *in silico* CGF fingerprints were determined for the 60 *L. monocytogenes* strains. A hierarchically clustered heatmap of the CGF fingerprints is shown here. Clusters defined at 20 SNPs (see Figure 3.1) are highlighted. Hierarchical clustering was performed on the the fingerprint similarity coefficients using the average linkage method (UPGMA).

3.3.2.2 *In silico* MLST

In silico MLST (Salcedo et al., 2003) data was generated for all 60 *L. monocytogenes* strains (Table 7). Sequence type (ST) designations were assigned according to the *L. monocytogenes* MLST database. These ST designations were used as the clusters derived from MLST typing and compared to the clusters derived from the other typing methods.

The *dat* locus could not be determined for H7858 since the locus was found to be truncated by the end of a contig. Alleles 2, 3 and 50 of the *dat* locus produced 100% ID full length matches to the 336 bp *dat* locus fragment retrieved from H7858. This match corresponded to the 136-471 bp region of the *dat* locus. Alleles 2, 3 and 50 of the *dat* locus varied by only 1 SNP in the 1-135 bp region of the allele. For the purposes of continuing the analysis with no missing data, the *dat* locus was assumed to be allele 3 for H7858, thus, the ST designation was assumed to be ST-6 as this was a more common allele in the *L. monocytogenes* MLST database. This ST designation was unique to H7858 in the set of 60 *L. monocytogenes* strains.

The allele designation for the *dat* MLST locus could likely not be retrieved for H7858 due to fragmentation of the whole genome sequence assembly. The assembly for H7858 was the most highly fragmented assembly of the 60 *L. monocytogenes* genome assemblies with 181 contigs. The mean and median number of contigs in the 60 genome assemblies was 15.43 and 1, respectively. More advanced sequence assemblers commonly used with short-read sequence data could possibly assemble the genome of H7858 if the read data were available for this genome. However, poor sequencing coverage may also prohibit assembly into fewer contigs.

3.3.2.3 *In silico* MVLST

In silico MVLST (Knabel et al., 2012; Zhang et al., 2004) data was generated for the 60 *L. monocytogenes* strains (Table 8). A database of alleles for each MVLST locus was constructed from the sequences obtained for each MVLST by Knabel et al. (2012). This database of alleles was expanded to include novel alleles present in the 60 *L. monocytogenes* genomes. MVLST ST were assigned in an incremental fashion starting at ST-1 and incrementing with each unique MVLST profile encountered in the dataset.

There were several instances where an allele match could not be found for the *inlC* MVLST locus. In F6854 and H7858, the allele designation could not be determined due to truncation of the *inlC* locus at the end of a contig. F6854 and H7858 contained the highest number of contigs of the 60 genome assemblies with 181 and 133 contigs, respectively. In all four lineage III strains (L99-4a, M7, HCC23, J2-071) in the set of 60 *L. monocytogenes* strains, the *inlC* locus was absent as signified by low nucleotide BLAST alignment coverage (21-45/416 bp; 95-78% ID). These data suggest that this *inlC* virulence gene locus may only be present within lineage I and II strains and not lineage III strains. The *inlC* gene codes for an internalin protein that may be required for virulence in certain mammilian host cell types. Since the majority of lineage III *L. monocytogenes* strains are non-pathogenic, this gene may have been dropped from the genome due to selective pressures favouring the loss of this gene.

3.3.2.4 Typing method performance assessment

Various metrics have been developed for the comparison and assessment of typing methods. One of the most commonly used metrics is the Simpson's index of diversity (D) (Simpson, 1949). The D when applied to typing methods provides a measure for determining the probability that two unrelated strains will be assigned to different typing groups (Hunter and Gaston, 1988). A typing method with a D of 0.0 would be expected to produce no diversity in typing profiles for a set of strains – all strains would be assigned to the same typing group – while a typing method with a D of 1.0 would be expected to assign each strain to its own typing group. Although the D has been a very useful measure of the discriminatory power of a typing method, it does not take into account the level of agreement between different typing methods (Carriço et al., 2006). Hence, in addition to determining the discriminatory power of a method, it is necessary to assess the concordance of the classifications assigned by the method to those assigned other typing methods.

When assessing the performance of one method respective to another, it is necessary to take into account both the discriminatory power of the method as well as how concordant the classifications derived from the method are to other methods. Carriço et al. (2006) have developed a framework for the comparison of typing methods through a variety of metrics to determine the discriminatory power of methods as well as the concordance between methods. To objectively compare the discriminatory power of typing methods, Carriço et al. (2006) suggest usage of D with confidence intervals (CI) as proposed by Grundmann et al. (2001). CI are used to account for variations in

the sample taken from the population and the sample size (Pinto et al., 2008; Severiano et al., 2011a). For assessing concordance between clusters derived from different methods, Carriço et al. (2006) suggest usage of the Adjusted Rand (AR) (Hubert and Arabie, 1985; Rand, 1971), Wallace (W) (Wallace, 1983) and Adjusted Wallace (AW) (Severiano et al., 2011b) coefficients. The Rand coefficient (R) can be used to determine typing method concordance (Versalovic et al., 1993). However, the R may overestimate concordance since it does not take into account the possibility of two clusters being generated by chance alone (Carriço et al., 2006). To correct for chance agreement, the AR was developed by Hubert and Arabie (1985) and provides a more accurate representation of the concordance between two methods. Unlike the Rand or AR coefficients, the W provides directional information between two typing methods (i.e., $W_{A\to B}$ and $W_{B\to A}$) (Wallace, 1983). This means that if two strains are found in the same cluster by one method, the W can tell us the probability that these two strains will be in the same cluster using another method. In order to increase the statistical support of comparison between W coefficients, Pinto et al. (2008) proposed the use of CI for the W and estimation of the expected W if the classifications were independent (W_i). Determination of a CI for the W of two classifications can help account for variability of clusters derived from different sampling of individuals and different sample sizes. Calculation of the W_i can determine if a high W can be explained by chance, or if a low W may actually be much higher. If, for example, when comparing two typing methods and the value of W_i is within the the CI of W, the typing methods likely produce clusters that are independent of one another and any cluster agreement likely occurred by chance (Pinto et al., 2008; Severiano et al., 2011b). Severiano et al. (2011b) proposed the AW as a correction to the W by directly taking the W_i into account. These metrics allow the comparison of methods on the basis of discriminatory power and concordance between clusters from one method compared to others.

When determining if a typing method has better performance than another method, it may be insufficient to only determine the concordance between classifications derived from each method. If method A has a high W respective to method B and method B has a low W respective to method A (i.e., $W_{A\to B} > W_{B\to A}$), all that is being established is that method A is better at inferring the classifications of B than method B is at inferring the classifications of A. However, these classifications may not be concordant with the phylogeny or epidemiology of the organism. To address this, our research group has proposed the usage of whole-genome phylogenies as the point of reference for the comparison of typing methods (Carrillo et al., 2012). Instead of

comparing one method against another, each method is compared against the whole-genome phylogeny. If the classifications from typing method A and the whole-genome phylogeny show better agreement than the classifications from typing method B and the whole-genome phylogeny (i.e., $W_{A\rightarrow\text{phylogeny}} > W_{B\rightarrow\text{phylogeny}}$), then typing method A could be considered to be the more phylogenetically informative method. Hence, by using whole-genome phylogenies as an external point of reference in the comparison of typing methods, a more accurate assessment of the performance of typing methods can be obtained.

The performance of the proposed CGF40 assay as well as the MLST and MVLST schemes for *L. monocytogenes* were assessed by determination of each method's concordance with respect to a whole-genome phylogeny. The MLST and MVLST schemes for *L. monocytogenes* were designed based on the assumption that one could subtype *L. monocytogenes* based on the sequences of several housekeeping or virulence gene loci. Genome sequences for multiple strains of *L. monocytogenes* were not available during the development of these typing schemes. However, with the advent of NGS, multiple strains from the major lineages of *L. monocytogenes* have been sequenced, and much of this WGS data has been made publicly available. Using this WGS data, we were able to determine potential markers for a *L. monocytogenes* CGF assay and select the markers that maximized concordance with PGC to produce a phylogenetically-informative and discriminatory genotyping assay. In order to determine the performance of this CGF assays constructed through optimized marker selection, we used an assessment framework using the D to determine the discriminatory power of the method and the AW to determine concordance to PGC, which are clusters derived from a whole-genome phylogeny.

The D was calculated for the PGC and *in silico* typing of the 60 *L. monocytogenes* strains (Table 3.1). The CGF40 assay with clusters defined at 100% fingerprint similarity produced the highest D (0.940). The PGC defined at 150 and 20 core genome SNPs produced D values of 0.874 and 0.901, respectively. The D of MLST and MVLST were 0.927 and 0.876, respectively. Only the D of the CGF40 assay at 100% was significantly greater than the D of the PGC at 20 SNPs ($P < 0.05$). The D of the CGF40 at 100% was significantly greater than MVLST ($P < 0.05$), but not MLST ($P \geq 0.05$). Hence, the CGF40 has greater discriminatory power than MVLST and the PGC defined at 20 core genome SNPs, but not MLST.

The D of MLST was significantly greater than the D of MVLST ($P < 0.05$) suggesting that MLST is a more discriminatory method than MVLST. This is contrary to the findings of Zhang

et al. (2004) where MVLST was found to have greater discriminatory power for serotype 1/2a and 4b strains than MLST. The high D of MLST may be explained by the sub-classification of the highly clonal group of serotype 1/2a Canadian outbreak-related strains (the blue phylogenomic cluster in Figures 3.2 and 3.1). The strains in this cluster have fewer than 20 SNPs between them. The MLST profiles of these strains differ by only a single locus (*abcZ*). Due to this single locus variation, MLST is able to sub-divide this phylogenomic cluster explaining the higher D for MLST compared to MVLST.

The AW coefficients with jackknife pseudo-values 95% CI were determined for MLST, MVLST and CGF40 clustering of the 60 *L. monocytogenes* strains based on *in silico* typing data for MLST, MVLST and CGF40 against PGC defined at 150 and 20 SNPs (Table 3.1). In order to determine if there was a significant difference ($P \leq 0.05$) between AW for different typing methods, p-values were calculated using the jackknife pseudo-values resampling method (Severiano et al., 2011a) (Table 3.2). The jackknife pseudo-values resampling method was used instead of the bootstrap method for calculating CI since it was shown by Severiano et al. (2011a) to more accurately estimate CI for pairwise agreement measures.

TABLE 3.1: The Adjusted Wallace (AW) with jackknife pseudo-values 95% confidence intervals (CI) are shown here for MLST, MVLST and CGF40 at various fingerprint similarity thresholds compared to the phylogenomic clusters (PGC) defined at 150 and 20 SNPs. The Simpson's index of diversity (D) with jackknife pseudo-values 95% CI are shown here for each clustering of the 60 *L. monocytogenes* strains. Typing data for each typing method was generated using MIST. AW and D with 95% CI were calculated using the online tool ComparingPartitions.

				150 SNPs		20 SNPs	
Method	Partitions	D	CI	AW	CI	AW	CI
CGF40; 100%	35	0.940	(0.892-0.987)	1.000	(1.000-1.000)	1.000	(1.000-1.000)
CGF40; 97.5%	31	0.929	(0.878-0.981)	1.000	(1.000-1.000)	1.000	(1.000-1.000)
CGF40; 95%	29	0.927	(0.876-0.978)	1.000	(1.000-1.000)	0.957	(0.877-1.000)
CGF40; 90%	22	0.877	(0.809-0.945)	0.990	(0.980-1.000)	0.776	(0.565-1.000)
MLST	22	0.927	(0.895-0.958)	1.000	(1.000-1.000)	0.710	(0.478-0.945)
MVLST	22	0.876	(0.808-0.944)	0.979	(0.954-1.000)	0.772	(0.559-1.000)
PGC; 20 SNPs	27	0.901	(0.835-0.966)				
PGC; 150 SNPs	19	0.874	(0.807-0.941)				

All typing methods were highly concordant with the PGC defined at 150 SNPs. Each method was able to produce similar or finer clusters than the PGC at 150 SNPs. However, with PGC defined at 20 SNPs, CGF40 was better able to produce similar or finer clusters than MLST or MVLST. *L. monocytogenes* CGF40 clusters at 97.5% and 100% produced significantly different ($P \leq 0.05$) AW compared to MLST ($P = 0.013$) and MVLST ($P = 0.046$). This indicates that

TABLE 3.2: The P between AW of different typing methods to the PGC defined at 20 SNPs (Table 3.1) are shown here. P were calculated using the jackknife pseudo-values resampling method with the online tool ComparingPartitions. $P \leq 0.05$ are highlighted.

$AW_{\text{(Method A→20 SNP clusters)}}$			$AW_{\text{(Method B→20 SNP clusters)}}$			
Method A	AW	CI	Method B	AW	CI	P
MLST	0.710	(0.478-0.945)	MVLST	0.772	(0.559-1.000)	0.197
CGF40; 95%	0.957	(0.877-1.000)	MVLST	0.772	(0.559-1.000)	0.057
CGF40; 97.5%	1.000	(1.000-1.000)	MVLST	0.772	(0.559-1.000)	**0.046**
CGF40; 100%	1.000	(1.000-1.000)	MVLST	0.772	(0.559-1.000)	**0.046**
CGF40; 95%	0.957	(0.877-1.000)	MLST	0.710	(0.478-0.945)	**0.008**
CGF40; 97.5%	1.000	(1.000-1.000)	MLST	0.710	(0.478-0.945)	**0.013**
CGF40; 100%	1.000	(1.000-1.000)	MLST	0.710	(0.478-0.945)	**0.013**

CGF40 at 97.5% and 100% fingerprint similarity was more concordant with the PGC at 20 SNPs than either MLST or MVLST ($P \leq 0.05$).

There was no significant difference between the AW for MLST and MVLST ($P \leq 0.05$). This suggests that there is no difference between the ability of MLST and MVLST to infer the PGC at 20 SNPs.

The proposed *L. monocytogenes* CGF40 assay designed using *in silico* optimized marker selection is a highly discriminatory assay with high concordance with the whole-genome phylogeny of 60 *L. monocytogenes* strains. The CGF40 assay has greater discriminatory power than MVLST and greater concordance with PGC defined at 20 SNPs than MLST or MVLST.

Chapter 4

Thesis Conclusions

As more bacterial strains are sequenced, we will be better able to assess the performance of current molecular typing methods and create new high-performance methods. Most conventional molecular typing methods were conceived prior to the genomics era. Comparative genomic studies have shown that, in some cases, molecular typing may inaccurately estimate the genetic similarity between bacterial isolates (Taboada et al., 2008). Furthermore, current molecular typing methods may fail to differentiate bacterial isolates in an epidemiologically and phylogenetically consistent manner; therefore, new methods may be required to distinguish these isolates (Xiao et al., 2011). Additionally, some "gold-standard" methods for typing of bacterial pathogens, such as PFGE for *L. monocytogenes* (Graves and Swaminathan, 2001; Swaminathan and Gerner-Smidt, 2007), are labour-intensive and time-consuming (Goering, 2010; Van Belkum et al., 2007). Although WGS may one day be the only molecular method for characterization of bacterial isolates for outbreak investigations for certain organisms (Gardy et al., 2011; Gilmour et al., 2010; Köser et al., 2012; Reimer et al., 2011; Rohde et al., 2011), for the foreseeable future, WGS will likely remain economically impractical for epidemiological surveillance. Therefore, inexpensive, high-throughput, informative and discriminatory molecular typing methods will be required for routine epidemiological surveillance of priority human bacterial pathogens.

Though *L. monocytogenes* has a high level of genomic synteny between strains from the major lineages (Deng et al., 2010; Hain et al., 2012), it has significant diversity in pan-genomic gene content and exhibits biased distribution of accessory genes across the major lineages (Deng et al.,

2010). The *Listeria* species have been reported to have the highest synonymous mutation rates of all prokaryotes following comparison of multiple completed genome sequences (Novichkov et al., 2009). Four lineages (I, II, III, IV (or IIIB)) and thirteen serotypes have been identified in *L. monocytogenes* (Piffaretti et al., 1989; Rasmussen et al., 1995; Ward et al., 2008; Wiedmann et al., 1997). Strains from serotypes 1/2a, 1/2b and 4b from lineages I and II represent the majority of human clinical cases of listeriosis (Farber and Peterkin, 1991). Serotypes 4a and 4c from lineages III and IV are commonly isolated from food, animals and the environment, but are rarely associated with human cases (Wiedmann et al., 1997). This is likely due to the absence of virulence factors in lineage III strains compared to strains from lineages I and II (Hain et al., 2012).

Our research group has proposed that typing methods be assessed on the basis of discriminatory power as well as concordance with whole-genome phylogenies (Carrillo et al., 2012). In this thesis, we incorporate and extend this idea in an *in silico* comparative genomic framework for the design and assessment of typing methods using WGS data. Using this framework, we were able to design an optimized CGF assay for *L. monocytogenes* with greater discriminatory power and concordance to a whole-genome phylogeny than the current typing methods, MLST and MVLST. Based on the work presented in this thesis, the following conclusions can be reached:

1. *In silico* typing can be used to generate multiple layers of typing data from WGS data, thereby, facilitating the comparison of typing methods.

2. Defining PGC can enable comparison of typing methods to WGS data through calculation of partition congruence metrics such as the Wallace and Adjusted Wallace coefficients.

3. In addition to assessing the performance of current typing methods, optimized marker selection with an *in silico* assessment framework as described in this thesis can be used to produce novel molecular typing methods that have the potential to outperform current molecular typing methods in terms of discriminatory power and concordance to whole-genome phylogenies.

References

Aarts, H. J., van Lith, L. A., and Jacobs-Reitsma, W. F. (1995). Discrepancy between penner serotyping and polymerase chain reaction fingerprinting of *Campylobacter* isolated from poultry and other animal sources. *Letters in Applied Microbiology*, 20(6):371–374. PMID: 7786504.

Achtman, M. and Wagner, M. (2008). Microbial diversity and the genetic nature of microbial species. *Nature Reviews. Microbiology*, 6(6):431–440. PMID: 18461076.

Achtman, M., Wain, J., Weill, Fran c.-X., Nair, S., Zhou, Z., Sangal, V., Krauland, M. G., Hale, J. L., Harbottle, H., Uesbeck, A., Dougan, G., Harrison, L. H., Brisse, S., and the S. enterica MLST study group (2012). Multilocus sequence typing as a replacement for serotyping in salmonella enterica. *PLoS Pathogens*, 8(6):e1002776.

Aeschbacher, M. and Piffaretti, J. C. (1989). Population genetics of human and animal enteric *Campylobacter* strains. *Infection and Immunity*, 57(5):1432–1437.

Akopyanz, N., Bukanov, N. O., Westblom, T. U., Kresovich, S., and Berg, D. E. (1992). DNA diversity among clinical isolates of *Helicobacter pylori* detected by PCR-based RAPD fingerprinting. *Nucleic Acids Research*, 20(19):5137–5142.

Alcaraz, L. D., Moreno-Hagelsieb, G., Eguiarte, L. E., Souza, V., Herrera-Estrella, L., and Olmedo, G. (2010). Understanding the evolutionary relationships and major traits of *Bacillus* through comparative genomics. *BMC Genomics*, 11:332. PMID: 20504335.

Alexander, D. C., Hao, W., Gilmour, M. W., Zittermann, S., Sarabia, A., Melano, R. G., Peralta, A., Lombos, M., Warren, K., Amatnieks, Y., Virey, E., Ma, J. H., Jamieson, F. B., Low, D. E., and Allen, V. G. (2012). *Escherichia coli* O104:H4 infections and international travel. *Emerging Infectious Diseases*, 18(3):473–476. PMID: 22377016.

Alm, R. A., Ling, L.-S. L., Moir, D. T., King, B. L., Brown, E. D., Doig, P. C., Smith, D. R., Noonan, B., Guild, B. C., deJonge, B. L., Carmel, G., Tummino, P. J., Caruso, A., Uria-Nickelsen, M., Mills, D. M., Ives, C., Gibson, R., Merberg, D., Mills, S. D., Jiang, Q., Taylor, D. E., Vovis, G. F., and Trust, T. J. (1999). Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen *Helicobacter pylori*. *Nature*, 397(6715):176–180.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410.

Atherton, J. C. (1998). *H. pylori* virulence factors. *British Medical Bulletin*, 54(1):105 –120.

Audurier, A. and Martin, C. (1989). Phage typing of *Listeria monocytogenes*. *International Journal of Food Microbiology*, 8(3):251–257.

Aziz, R. K., Bartels, D., Best, A. A., DeJongh, M., Disz, T., Edwards, R. A., Formsma, K., Gerdes, S., Glass, E. M., Kubal, M., Meyer, F., Olsen, G. J., Olson, R., Osterman, A. L., Overbeek, R. A., McNeil, L. K., Paarmann, D., Paczian, T., Parrello, B., Pusch, G. D., Reich, C., Stevens, R., Vassieva, O., Vonstein, V., Wilke, A., and Zagnitko, O. (2008). The RAST server: rapid annotations using subsystems technology. *BMC Genomics*, 9:75. PMID: 18261238.

Baggesen, D. L., Sørensen, G., Nielsen, E. M., and Wegener, H. C. (2010). Phage typing of *Salmonella typhimurium* - is it still a useful tool for surveillance and outbreak investigation? *Euro Surveillance: Bulletin Europen Sur Les Maladies Transmissibles = European Communicable Disease Bulletin*, 15(4):19471. PMID: 20122382.

Balandyt, L., Brodard, I., Frey, J., Oevermann, A., and Abril, C. (2011). Ruminant rhombencephalitis-associated *Listeria monocytogenes* alleles linked to a multilocus variable-number tandem-repeat analysis complex. *Applied and Environmental Microbiology*, 77(23):8325–8335.

Baumler, D. J., Peplinski, R. G., Reed, J. L., Glasner, J. D., and Perna, N. T. (2011). The evolution of metabolic networks of *E. coli*. *BMC Systems Biology*, 5(1):182.

Behringer, M., Miller, W. G., and Oyarzabal, O. A. (2011). Typing of *Campylobacter jejuni* and *Campylobacter coli* isolated from live broilers and retail broiler meat by flaA-RFLP, MLST, PFGE and REP-PCR. *Journal of Microbiological Methods*, 84(2):194–201. PMID: 21130125.

Bennett, S. (2004). Solexa ltd. *Pharmacogenomics*, 5(4):433–438.

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., Hall, K. P., Evers, D. J., Barnes, C. L., Bignell, H. R., Boutell, J. M., Bryant, J., Carter, R. J., Cheetham, R. K., Cox, A. J., Ellis, D. J., Flatbush, M. R., Gormley, N. A., Humphray, S. J., Irving, L. J., Karbelashvili, M. S., Kirk, S. M., Li, H., Liu, X., Maisinger, K. S., Murray, L. J., Obradovic, B., Ost, T., Parkinson, M. L., Pratt, M. R., Rasolonjatovo, I. M. J., Reed, M. T., Rigatti, R., Rodighiero, C., Ross, M. T., Sabot, A., Sankar, S. V., Scally, A., Schroth, G. P., Smith, M. E., Smith, V. P., Spiridou, A., Torrance, P. E., Tzonev, S. S., Vermaas, E. H., Walter, K., Wu, X., Zhang, L., Alam, M. D., Anastasi, C., Aniebo, I. C., Bailey, D. M. D., Bancarz, I. R., Banerjee, S., Barbour, S. G., Baybayan, P. A., Benoit, V. A., Benson, K. F., Bevis, C., Black, P. J., Boodhun, A., Brennan, J. S., Bridgham, J. A., Brown, R. C., Brown, A. A., Buermann, D. H., Bundu, A. A., Burrows, J. C., Carter, N. P., Castillo, N., Catenazzi, M. C. E., Chang, S., Cooley, R. N., Crake, N. R., Dada, O. O., Diakoumakos, K. D., Dominguez-Fernandez, B., Earnshaw, D. J., Egbujor, U. C., Elmore, D. W., Etchin, S. S., Ewan, M. R., Fedurco, M., Fraser, L. J., Fajardo, K. V. F., Furey, W. S., George, D., Gietzen, K. J., Goddard, C. P., Golda, G. S., Granieri, P. A., Green, D. E., Gustafson, D. L., Hansen, N. F., Harnish, K., Haudenschild, C. D., Heyer, N. I., Hims, M. M., Ho, J. T., Horgan, A. M., Hoschler, K., Hurwitz, S., Ivanov, D. V., Johnson, M. Q., James, T., Jones, T. A. H., Kang, G.-D., Kerelska, T. H., Kersey, A. D., Khrebtukova, I., Kindwall, A. P., Kingsbury, Z., Kokko-Gonzales, P. I., Kumar, A., Laurent, M. A., Lawley, C. T., Lee, S. E., Lee, X., Liao, A. K., Loch, J. A., Lok, M., Luo, S., Mammen, R. M., Martin, J. W., McCauley, P. G., McNitt, P., Mehta, P., Moon, K. W., Mullens, J. W., Newington, T., Ning, Z., Ng, B. L., Novo, S. M., O'Neill, M. J., Osborne, M. A., Osnowski, A., Ostadan, O., Paraschos, L. L., Pickering, L., Pike, A. C., Pike, A. C., Pinkard, D. C., Pliskin, D. P., Podhasky, J., Quijano, V. J., Raczy, C., Rae, V. H., Rawlings, S. R., Rodriguez, A. C., Roe, P. M., Rogers, J., Bacigalupo, M. C. R., Romanov, N., Romieu, A., Roth, R. K., Rourke, N. J., Ruediger, S. T., Rusman, E., Sanches-Kuiper, R. M., Schenker, M. R., Seoane, J. M., Shaw, R. J., Shiver, M. K., Short, S. W., Sizto, N. L., Sluis, J. P., Smith, M. A., Sohna, J. E. S., Spence, E. J., Stevens, K., Sutton, N., Szajkowski, L., Tregidgo, C. L., Turcatti, G., vandeVondele, S., Verhovsky, Y., Virk, S. M., Wakelin, S., Walcott, G. C., Wang, J., Worsley, G. J., Yan, J., Yau, L., Zuerlein, M., Rogers, J., Mullikin, J. C., Hurles, M. E., McCooke, N. J., West, J. S., Oaks, F. L., Lundberg, P. L., Klenerman, D., Durbin, R., and Smith, A. J. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456(7218):53–59.

Bochner, B. R., Gadzinski, P., and Panomitros, E. (2001). Phenotype microarrays for high-throughput phenotypic testing and assay of gene function. *Genome Research*, 11(7):1246–1255.

Boerlin, P., Bannerman, E., Jemmi, T., and Bille, J. (1996). Subtyping *Listeria monocytogenes* isolates genetically related to the Swiss epidemic clone. *Journal of Clinical Microbiology*, 34(9):2148–2153. PMID: 8862575.

Cai, S., Kabuki, D. Y., Kuaye, A. Y., Cargioli, T. G., Chung, M. S., Nielsen, R., and Wiedmann, M. (2002). Rational design of DNA sequence-based strategies for subtyping *Listeria monocytogenes*. *Journal of Clinical Microbiology*, 40(9):3319–3325. PMID: 12202573 PMCID: PMC130781.

Call, D. R., Borucki, M. K., and Besser, T. E. (2003). Mixed-genome microarrays reveal multiple serotype and lineage-specific differences among strains of *Listeria monocytogenes*. *Journal of Clinical Microbiology*, 41(2):632–639.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T. L. (2009). BLAST+: architecture and applications. *BMC Bioinformatics*, 10:421. PMID: 20003500.

Campioni, F., Moratto Bergamini, A. M., and Falcão, J. P. (2012). Genetic diversity, virulence genes and antimicrobial resistance of *Salmonella* enteritidis isolated from food and humans over a 24-year period in Brazil. *Food Microbiology*, 32(2):254–264. PMID: 22986188.

Carriço, J. A., Silva-Costa, C., Melo-Cristino, J., Pinto, F. R., de Lencastre, H., Almeida, J. S., and Ramirez, M. (2006). Illustration of a common framework for relating multiple typing methods by application to macrolide-resistant *Streptococcus pyogenes*. *Journal of Clinical Microbiology*, 44(7):2524–2532.

Carrillo, C. D., Kruczkiewicz, P., Mutschall, S., Tudor, A., Clark, C., and Taboada, E. N. (2012). A framework for assessing the concordance of molecular typing methods and the true strain phylogeny of *Campylobacter jejuni* and *C. coli* using draft genome sequence data. *Frontiers in Cellular and Infection Microbiology*, 2:57. PMID: 22919648 PMCID: PMC3417556.

Chan, M. S., Maiden, M. C., and Spratt, B. G. (2001). Database-driven multi locus sequence typing (MLST) of bacterial pathogens. *Bioinformatics*, 17(11):1077–1083. PMID: 11724739.

Chizhikov, V., Rasooly, A., Chumakov, K., and Levy, D. D. (2001). Microarray analysis of microbial virulence factors. *Applied and Environmental Microbiology*, 67(7):3258–3263. PMID: 11425749.

Clark, C. G., Taboada, E., Grant, C. C. R., Blakeston, C., Pollari, F., Marshall, B., Rahn, K., MacKinnon, J., Daignault, D., Pillai, D., and Ng, L.-K. (2012). Comparison of molecular typing methods useful for detecting clusters of *Campylobacter jejuni* and *C. coli* isolates through routine surveillance. *Journal of Clinical Microbiology*, 50(3):798–809.

Comas, I., Homolka, S., Niemann, S., and Gagneux, S. (2009). Genotyping of genetically monomorphic bacteria: DNA sequencing in *Mycobacterium tuberculosis* highlights the limitations of current methodologies. *PLoS ONE*, 4(11):e7815.

Conly, J. and Johnston, B. (2008). Listeria: A persistent food-borne pathogen. *The Canadian Journal of Infectious Diseases & Medical Microbiology*, 19(5):327–328. PMID: 19436470 PMCID: PMC2606629.

D'Agata, E. M., Gerrits, M. M., Tang, Y., Samore, M., and Kusters, J. G. (2001). Comparison of pulsedfield gel electrophoresis and amplified fragmentlength polymorphism for epidemiological investigations of common nosocomial pathogens. *Infection Control and Hospital Epidemiology*, 22(9):550–554.

Dass, S. C., Abu-Ghannam, N., Antony-Babu, S., and Cummins, E. J. (2010). Ecology and molecular typing of *L. monocytogenes* in a processing plant for cold-smoked salmon in the *Republic of Ireland*. *Food Research International*, 43(5):1529–1536.

den Bakker, H. C., Bundrant, B. N., Fortes, E. D., Orsi, R. H., and Wiedmann, M. (2010). A population genetics-based and phylogenetic approach to understanding the evolution of virulence in the genus *Listeria*. *Applied and Environmental Microbiology*, 76(18):6085–6100. PMID: 20656873.

Deng, X., Phillippy, A. M., Li, Z., Salzberg, S. L., and Zhang, W. (2010). Probing the pangenome of *Listeria monocytogenes*: new insights into intraspecific niche expansion and genomic diversification. *BMC Genomics*, 11(1):500.

Dijkshoorn, L., Aucken, H., Gerner-Smidt, P., Janssen, P., Kaufmann, M. E., Garaizar, J., Ursing, J., and Pitt, T. L. (1996). Comparison of outbreak and nonoutbreak *Acinetobacter*

baumannii strains by genotypic and phenotypic methods. *Journal of Clinical Microbiology*, 34(6):1519–1525. PMID: 8735109 PMCID: PMC229053.

Dorrell, N., Mangan, J. A., Laing, K. G., Hinds, J., Linton, D., Al-Ghusein, H., Barrell, B. G., Parkhill, J., Stoker, N. G., Karlyshev, A. V., Butcher, P. D., and Wren, B. W. (2001). Whole genome comparison of *Campylobacter jejuni* human isolates using a low-cost microarray reveals extensive genetic diversity. *Genome Research*, 11(10):1706–1715. PMID: 11591647 PMCID: PMC311159.

Doumith, M., Buchrieser, C., Glaser, P., Jacquet, C., and Martin, P. (2004a). Differentiation of the major *Listeria monocytogenes* serovars by multiplex PCR. *Journal of Clinical Microbiology*, 42(8):3819–3822.

Doumith, M., Cazalet, C., Simoes, N., Frangeul, L., Jacquet, C., Kunst, F., Martin, P., Cossart, P., Glaser, P., and Buchrieser, C. (2004b). New aspects regarding evolution and virulence of *Listeria monocytogenes* revealed by comparative genomics and DNA arrays. *Infection and Immunity*, 72(2):1072–1083. PMID: 14742555 PMCID: PMC321639.

Doumith, M., Jacquet, C., Goulet, V., Oggioni, C., Van Loock, F., Buchrieser, C., and Martin, P. (2006). Use of DNA arrays for the analysis of outbreak-related strains of *Listeria monocytogenes*. *International Journal of Medical Microbiology: IJMM*, 296(8):559–562. PMID: 17002895.

Ducey, T. F., Page, B., Usgaard, T., Borucki, M. K., Pupedis, K., and Ward, T. J. (2007). A single-nucleotide-polymorphism-based multilocus genotyping assay for subtyping lineage i isolates of *Listeria monocytogenes*. *Applied and Environmental Microbiology*, 73(1):133–147. PMID: 17085705.

Duong, T. and Konkel, M. E. (2009). Comparative studies of *Campylobacter jejuni* genomic diversity reveal the importance of core and dispensable genes in the biology of this enigmatic food-borne pathogen. *Current Opinion in Biotechnology*, 20(2):158–165. PMID: 19346123.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5):1792 –1797.

Edman, D. C., Pollock, M. B., and Hall, E. R. (1968). Listeria monocytogenes l forms i. induction, maintenance, and biological characteristics. *Journal of Bacteriology*, 96(2):352–357.

Farber, J. M. and Peterkin, P. I. (1991). *Listeria monocytogenes*, a food-borne pathogen. *Microbiological Reviews*, 55(3):476–511. PMID: 1943998 PMCID: PMC372831.

Feil, E. J., Li, B. C., Aanensen, D. M., Hanage, W. P., and Spratt, B. G. (2004). eBURST: inferring patterns of evolutionary descent among clusters of related bacterial genotypes from multilocus sequence typing data. *Journal of Bacteriology*, 186(5):1518–1530.

Finney, M. (2001). Pulsed-field gel electrophoresis. *Current Protocols in Molecular Biology / Edited By Frederick M. Ausubel ... [et Al.]*, Chapter 2:Unit2.5B. PMID: 18265186.

Fouts, D. E., Mongodin, E. F., Mandrell, R. E., Miller, W. G., Rasko, D. A., Ravel, J., Brinkac, L. M., DeBoy, R. T., Parker, C. T., Daugherty, S. C., Dodson, R. J., Durkin, A. S., Madupu, R., Sullivan, S. A., Shetty, J. U., Ayodeji, M. A., Shvartsbeyn, A., Schatz, M. C., Badger, J. H., Fraser, C. M., and Nelson, K. E. (2005). Major structural differences and novel potential virulence mechanisms from the genomes of multiple *Campylobacter* species. *PLoS Biology*, 3(1):e15. PMID: 15660156.

Foxman, B. and Riley, L. (2001). Molecular epidemiology: Focus on infection. *American Journal of Epidemiology*, 153(12):1135 –1141.

Francisco, A., Bugalho, M., Ramirez, M., and Carrio, J. (2009). Global optimal eBURST analysis of multilocus typing data using a graphic matroid approach. *BMC Bioinformatics*, 10(1):152.

Franklin, K., Lingohr, E. J., Yoshida, C., Anjum, M., Bodrossy, L., Clark, C. G., Kropinski, A. M., and Karmali, M. A. (2011). Rapid genoserotyping tool for classification of *Salmonella* serovars. *Journal of Clinical Microbiology*, 49(8):2954–2965. PMID: 21697324 PMCID: PMC3147765.

Gardy, J. L., Johnston, J. C., Ho Sui, S. J., Cook, V. J., Shah, L., Brodkin, E., Rempel, S., Moore, R., Zhao, Y., Holt, R., Varhol, R., Birol, I., Lem, M., Sharma, M. K., Elwood, K., Jones, S. J. M., Brinkman, F. S. L., Brunham, R. C., and Tang, P. (2011). Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. *The New England Journal of Medicine*, 364(8):730–739. PMID: 21345102.

Gilmour, M. W., Graham, M., Domselaar, G. V., Tyler, S., Kent, H., Trout-Yakel, K. M., Larios, O., Allen, V., Lee, B., and Nadon, C. (2010). High-throughput genome sequencing of two *Listeria monocytogenes* clinical isolates during a large foodborne outbreak. *BMC Genomics*, 11(1):120.

Glaser, P., Frangeul, L., Buchrieser, C., Rusniok, C., Amend, A., Baquero, F., Berche, P., Bloecker, H., Brandt, P., Chakraborty, T., Charbit, A., Chetouani, F., Couve, E., de Daruvar, A., Dehoux, P., Domann, E., Dominguez-Bernal, G., Duchaud, E., Durant, L., Dussurget, O., Entian, K.-D., Fsihi, H., Portillo, F. G.-D., Garrido, P., Gautier, L., Goebel, W., Gomez-Lopez, N., Hain, T., Hauf, J., Jackson, D., Jones, L.-M., Kaerst, U., Kreft, J., Kuhn, M., Kunst, F., Kurapkat, G., Madueno, E., Maitournam, A., Vicente, J. M., Ng, E., Nedjari, H., Nordsiek, G., Novella, S., de Pablos, B., Perez-Diaz, J.-C., Purcell, R., Remmel, B., Rose, M., Schlueter, T., Simoes, N., Tierrez, A., Vazquez-Boland, J.-A., Voss, H., Wehland, J., and Cossart, P. (2001). Comparative genomics of *Listeria* species. *Science*, 294(5543):849–852.

Goering, R. V. (2010). Pulsed field gel electrophoresis: A review of application and interpretation in the molecular epidemiology of infectious disease. *Infection, Genetics and Evolution*, 10(7):866–875.

Government of Canada, P. H. A. o. C. (2008). Link between listeriosis outbreak strain and maple leaf foods products confirmed. http://www.phac-aspc.gc.ca/media/nr-rp/2008/2008_13-eng.php. The Public Health Agency of Canada and the Canadian Food Inspection Agency have received laboratory results from Health Canada that establish a link between meat products recalled by Maple Leaf Foods from their plant in Toronto and an outbreak of listeriosis in four provinces. To date, 21 cases of listeriosis have been confirmed, and the same strain has been detected in four people who have died. A further 30 cases remain under investigation.

Government of Canada, P. H. A. o. C. (2009). Lessons learned: Public Health Agency of Canada'sresponse to the 2008 listeriosis outbreak - Public Health Agency of Canada. http://www.phac-aspc.gc.ca/fs-sa/listeria/2008-lessons-lecons-eng.php. Presents a general overview to the Agencys senior management of what worked well during the outbreak, and what needs further refinement, in order for the Agency to be better prepared for future outbreaks.

Graves, L. M. and Swaminathan, B. (2001). PulseNet standardized protocol for subtyping *Listeria monocytogenes* by macrorestriction and pulsed-field gel electrophoresis. *International Journal of Food Microbiology*, 65(1-2):55–62. PMID: 11322701.

Griffith, F. (1928). The significance of pneumococcal types. *Epidemiology & Infection*, 27(02):113–159.

Grundmann, H., Hori, S., and Tanner, G. (2001). Determining confidence intervals when measuring genetic diversity and the discriminatory abilities of typing methods for microorganisms. *Journal of Clinical Microbiology*, 39(11):4190–4192.

Hain, T., Ghai, R., Billion, A., Kuenne, C. T., Steinweg, C., Izar, B., Mohamed, W., Mraheil, M., Domann, E., Schaffrath, S., Kärst, U., Goesmann, A., Oehm, S., Phler, A., Merkl, R., Vorwerk, S., Glaser, P., Garrido, P., Rusniok, C., Buchrieser, C., Goebel, W., and Chakraborty, T. (2012). Comparative genomics and transcriptomics of lineages i, II, and III strains of *Listeria monocytogenes*. *BMC Genomics*, 13(1):144.

Hall, B. G., Ehrlich, G. D., and Hu, F. Z. (2010). Pan-genome analysis provides much higher strain typing resolution than multi-locus sequence typing. *Microbiology (Reading, England)*, 156(Pt 4):1060–1068. PMID: 20019077.

Hayashi, T., Makino, K., Ohnishi, M., Kurokawa, K., Ishii, K., Yokoyama, K., Han, C.-G., Ohtsubo, E., Nakayama, K., Murata, T., Tanaka, M., Tobe, T., Iida, T., Takami, H., Honda, T., Sasakawa, C., Ogasawara, N., Yasunaga, T., Kuhara, S., Shiba, T., Hattori, M., and Shinagawa, H. (2001). Complete genome sequence of enterohemorrhagic *Eschelichia coli* O157:H7 and genomic comparison with a laboratory strain K-12. *DNA Research*, 8(1):11 –22.

Howard-Jones, N. (1984). Robert Koch and the cholera vibrio: a centenary. *British Medical Journal (Clinical Research Ed.)*, 288(6414):379–381. PMID: 6419937 PMCID: PMC1444283.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of Classification*, 2(1):193–218.

Hunter, P. R. and Gaston, M. A. (1988). Numerical index of the discriminatory ability of typing systems: an application of simpson's index of diversity. *Journal of Clinical Microbiology*, 26(11):2465–2466.

Israel, D. A., Salama, N., Arnold, C. N., Moss, S. F., Ando, T., Wirth, H.-P., Tham, K. T., Camorlinga, M., Blaser, M. J., Falkow, S., and Peek, R. M. (2001). *Helicobacter pylori* strain-specific differences in genetic content, identified by microarray, influence host inflammatory responses. *Journal of Clinical Investigation*, 107(5):611–620.

Janssen, P., Maquelin, K., Coopman, R., Tjernberg, I., Bouvet, P., Kersters, K., and Dijkshoorn, L. (1997). Discrimination of *Acinetobacter* genomic species by AFLP fingerprinting. *International Journal of Systematic Bacteriology*, 47(4):1179–1187. PMID: 9336926.

Jolley, K. A., Chan, M.-S., and Maiden, M. C. (2004). mlstdbNet distributed multi-locus sequence typing (MLST) databases. *BMC Bioinformatics*, 5:86. PMID: 15230973 PMCID: PMC459212.

Kaplinski, L., Andreson, R., Puurand, T., and Remm, M. (2005). MultiPLX: automatic grouping and evaluation of PCR primers. *Bioinformatics*, 21(8):1701–1702.

Kato-Maeda, M., Rhee, J. T., Gingeras, T. R., Salamon, H., Drenkow, J., Smittipat, N., and Small, P. M. (2001). Comparing genomes within the species *Mycobacterium tuberculosis*. *Genome Research*, 11(4):547–554. PMID: 11282970.

Kauffmann, F. (1975). *Classification of bacteria: a realistic scheme with special reference to the classification of* Salmonella- *and* Escherichia-*species*. Munksgaard.

Kérouanton, A., Marault, M., Petit, L., Grout, J., Dao, T. T., and Brisabois, A. (2010). Evaluation of a multiplex PCR assay as an alternative method for *Listeria monocytogenes* serotyping. *Journal of Microbiological Methods*, 80(2):134–137.

Killgore, G., Thompson, A., Johnson, S., Brazier, J., Kuijper, E., Pepin, J., Frost, E. H., Savelkoul, P., Nicholson, B., Berg, R. J. v. d., Kato, H., Sambol, S. P., Zukowski, W., Woods, C., Limbago, B., Gerding, D. N., and McDonald, L. C. (2008). Comparison of seven techniques for typing international epidemic strains of *Clostridium difficile*: Restriction endonuclease analysis, pulsed-field gel electrophoresis, PCR-ribotyping, multilocus sequence typing, multilocus variable-number tandem-repeat analysis, amplified fragment length polymorphism, and surface layer protein A gene sequence typing. *Journal of Clinical Microbiology*, 46(2):431–437.

Knabel, S. J., Reimer, A., Verghese, B., Lok, M., Ziegler, J., Farber, J., Pagotto, F., Graham, M., Nadon, C. A., and Gilmour, M. W. (2012). Sequence typing confirms that a predominant *Listeria monocytogenes* clone caused human listeriosis cases and outbreaks in Canada from 1988 to 2010. *Journal of Clinical Microbiology*, 50(5):1748–1751. PMID: 22337989.

Köser, C. U., Holden, M. T. G., Ellington, M. J., Cartwright, E. J. P., Brown, N. M., Ogilvy-Stuart, A. L., Hsu, L. Y., Chewapreecha, C., Croucher, N. J., Harris, S. R., Sanders, M., Enright, M. C., Dougan, G., Bentley, S. D., Parkhill, J., Fraser, L. J., Betley, J. R., Schulz-Trieglaff, O. B., Smith, G. P., and Peacock, S. J. (2012). Rapid whole-genome sequencing for investigation of a neonatal MRSA outbreak. *The New England Journal of Medicine*, 366(24):2267–2275. PMID: 22693998.

Kruczkiewicz, P., Mutschall, S., Barker, D., Thomas, J., Van Domselaar, G., Gannon, V. P., Carrillo, C. D., and Taboada, E. N. (2013). MIST: a tool for rapid *in silico* generation of molecular data from bacterial genome sequences. *Proceedings of Bioinformatics 2013: 4th International Conference on Bioinformatics Models, Methods and Algorithms*, pages 316–323.

Kruczkiewicz, P., Tudor, A., Mutschall, S. K., Buchanan, C. J., Laing, C. R., Thomas, J. E., Gannon, V. P., Clark, C. G., Carrillo, C. D., and Taboada, E. N. (2011). A bioinformatics toolkit for comparative genomic analysis of *Campylobacter jejuni* in support of next-generation genotyping methods design. 16th International Workshop on Campylobacter, Helicobacter and Related Organisms (CHRO; poster).

Laing, C., Pegg, C., Yawney, D., Ziebell, K., Steele, M., Johnson, R., Thomas, J. E., Taboada, E. N., Zhang, Y., and Gannon, V. P. J. (2008). Rapid determination of *Escherichia coli* O157:H7 lineage types and molecular subtypes by using comparative genomic fingerprinting. *Appl. Environ. Microbiol.*, 74(21):6606–6615.

Laksanalamai, P., Joseph, L. A., Silk, B. J., Burall, L. S., L. Tarr, C., Gerner-Smidt, P., and Datta, A. R. (2012). Genomic characterization of *Listeria monocytogenes* strains involved in a multistate listeriosis outbreak associated with cantaloupe in US. *PLoS ONE*, 7(7):e42448.

Lancefield, R. C. (1933). A serological differentiation of human and other groups of hemolytic streptococci. *The Journal of Experimental Medicine*, 57(4):571–595.

Lashkari, D. A., DeRisi, J. L., McCusker, J. H., Namath, A. F., Gentile, C., Hwang, S. Y., Brown, P. O., and Davis, R. W. (1997). Yeast microarrays for genome wide parallel genetic and gene expression analysis. *Proceedings of the National Academy of Sciences of the United States of America*, 94(24):13057–13062. PMID: 9371799.

Leopold, S. R., Magrini, V., Holt, N. J., Shaikh, N., Mardis, E. R., Cagno, J., Ogura, Y., Iguchi, A., Hayashi, T., Mellmann, A., Karch, H., Besser, T. E., Sawyer, S. A., Whittam, T. S., and Tarr, P. I. (2009). A precise reconstruction of the emergence and constrained radiations of *Escherichia coli* O157 portrayed by backbone concatenomic analysis. *Proceedings of the National Academy of Sciences*, 106(21):8713 –8718.

Li, L., Stoeckert, C. J., and Roos, D. S. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Research*, 13(9):2178–2189. PMID: 12952885 PMCID: PMC403725.

Lindstedt, B.-A., Tham, W., Danielsson-Tham, M.-L., Vardund, T., Helmersson, S., and Kapperud, G. (2008). Multiple-locus variable-number tandem-repeats analysis of *Listeria monocytogenes* using multicolour capillary electrophoresis and comparison with pulsed-field gel electrophoresis typing. *Journal of Microbiological Methods*, 72(2):141–148. PMID: 18096258.

Lucchini, S., Thompson, A., and Hinton, J. C. D. (2001). Microarrays for microbiologists. *Microbiology*, 147(6):1403–1414.

Lukjancenko, O., Wassenaar, T. M., and Ussery, D. W. (2010). Comparison of 61 sequenced *Escherichia coli* genomes. *Microbial Ecology*. PMID: 20623278.

Maiden, M. C., Bygraves, J. A., Feil, E., Morelli, G., Russell, J. E., Urwin, R., Zhang, Q., Zhou, J., Zurth, K., Caugant, D. A., Feavers, I. M., Achtman, M., and Spratt, B. G. (1998). Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proceedings of the National Academy of Sciences of the United States of America*, 95(6):3140–3145. PMID: 9501229.

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y.-J., Chen, Z., Dewell, S. B., Du, L., Fierro, J. M., Gomes, X. V., Godwin, B. C., He, W., Helgesen, S., Ho, C. H., Irzyk, G. P., Jando, S. C., Alenquer, M. L. I., Jarvie, T. P., Jirage, K. B., Kim, J.-B., Knight, J. R., Lanza, J. R., Leamon, J. H., Lefkowitz, S. M., Lei, M., Li, J., Lohman, K. L., Lu, H., Makhijani, V. B., McDade, K. E., McKenna, M. P., Myers, E. W., Nickerson, E., Nobile, J. R., Plant, R., Puc, B. P., Ronan, M. T., Roth, G. T., Sarkis, G. J., Simons, J. F., Simpson, J. W., Srinivasan, M., Tartaro, K. R., Tomasz, A., Vogt, K. A., Volkmer, G. A., Wang, S. H., Wang, Y., Weiner, M. P., Yu, P., Begley, R. F., and Rothberg, J. M. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057):376–380.

McConnell, M. M., Thomas, L. V., Day, N. P., and Rowe, B. (1985). Enzyme-linked immunosorbent assays for the detection of adhesion factor antigens of enterotoxigenic *Escherichia coli*. *The Journal of Infectious Diseases*, 152(6):1120–1127. PMID: 3934290.

McLauchlin, J. (1990). Distribution of serovars of *Listeria monocytogenes* isolated from different categories of patients with listeriosis. *European Journal of Clinical Microbiology & Infectious Diseases: Official Publication of the European Society of Clinical Microbiology*, 9(3):210–213. PMID: 2110901.

Mead, P. S., Slutsker, L., Dietz, V., McCaig, L. F., Bresee, J. S., Shapiro, C., Griffin, P. M., and Tauxe, R. V. (1999). Food-related illness and death in the united states. *Emerging Infectious Diseases*, 5(5):607–625. PMID: 10511517.

Medini, D., Donati, C., Tettelin, H., Masignani, V., and Rappuoli, R. (2005). The microbial pan-genome. *Current Opinion in Genetics & Development*, 15(6):589–594. PMID: 16185861.

Merchant-Patel, S., Blackall, P. J., Templeton, J., Price, E. P., Miflin, J. K., Huygens, F., and Giffard, P. M. (2008). Characterisation of chicken *Campylobacter jejuni* isolates using resolution optimised single nucleotide polymorphisms and binary gene markers. *International Journal of Food Microbiology*, 128(2):304–308. PMID: 18835503.

Merrill, R. (2010). *Introduction to Epidemiology*. Jones & Bartlett Learning.

Miller, J. M. and Rhoden, D. L. (1991). Preliminary evaluation of Biolog, a carbon source utilization method for bacterial identification. *Journal of Clinical Microbiology*, 29(6):1143–1147.

Miya, S., Kimura, B., Sato, M., Takahashi, H., Ishikawa, T., Suda, T., Takakura, C., Fujii, T., and Wiedmann, M. (2008). Development of a multilocus variable-number of tandem repeat typing method for *Listeria monocytogenes* serotype 4b strains. *International Journal of Food Microbiology*, 124(3):239–249. PMID: 18457891.

Monot, M., Honoré, N., Garnier, T., Zidane, N., Sherafi, D., Paniz-Mondolfi, A., Matsuoka, M., Taylor, G. M., Donoghue, H. D., Bouwman, A., Mays, S., Watson, C., Lockwood, D., Khamesipour, A., Khamispour, A., Dowlati, Y., Jianping, S., Rea, T. H., Vera-Cabrera, L., Stefani, M. M., Banu, S., Macdonald, M., Sapkota, B. R., Spencer, J. S., Thomas, J., Harshman, K., Singh, P., Busso, P., Gattiker, A., Rougemont, J., Brennan, P. J., and Cole, S. T. (2009). Comparative genomic and phylogeographic analysis of *Mycobacterium leprae*. *Nature Genetics*, 41(12):1282–1289. PMID: 19881526.

Murphy, M., Corcoran, D., Buckley, J. F., O'Mahony, M., Whyte, P., and Fanning, S. (2007). Development and application of multiple-locus variable number of tandem repeat analysis (MLVA) to subtype a collection of *Listeria monocytogenes*. *International Journal of Food Microbiology*, 115(2):187–194. PMID: 17174430.

Musser, J. M., Amin, A., and Ramaswamy, S. (2000). Negligible genetic diversity of *Mycobacterium tuberculosis* host immune system protein targets: evidence of limited selective pressure. *Genetics*, 155(1):7–16. PMID: 10790380.

Nachamkin, I., Bohachick, K., and Patton, C. M. (1993). Flagellin gene typing of *Campylobacter jejuni* by restriction fragment length polymorphism analysis. *Journal of Clinical Microbiology*, 31(6):1531–1536. PMID: 8100241 PMCID: PMC265573.

Nightingale, K. K., Milillo, S. R., Ivy, R. A., Ho, A. J., Oliver, H. F., and Wiedmann, M. (2007). *Listeria monocytogenes* F2365 carries several authentic mutations potentially leading to truncated gene products, including inlB, and demonstrates atypical phenotypic characteristics. *Journal of Food Protection*, 70(2):482–488. PMID: 17340887.

Novichkov, P. S., Wolf, Y. I., Dubchak, I., and Koonin, E. V. (2009). Trends in prokaryotic evolution revealed by comparison of closely related bacterial and archaeal genomes. *Journal of Bacteriology*, 191(1):65–73. PMID: 18978059.

Odumeru, J. A., Steele, M., Fruhner, L., Larkin, C., Jiang, J., Mann, E., and McNab, W. B. (1999). Evaluation of accuracy and repeatability of identification of food-borne pathogens by automated bacterial identification systems. *Journal of Clinical Microbiology*, 37(4):944–949.

Olson, A. B., Andrysiak, A. K., Tracz, D. M., Guard-Bouldin, J., Demczuk, W., Ng, L.-K., Maki, A., Jamieson, F., and Gilmour, M. W. (2007). Limited genetic diversity in *Salmonella enterica* serovar Enteritidis PT13. *BMC Microbiology*, 7(1):87. PMID: 17908316.

Pagotto, F., Ng, L.-K., Clark, C., and Farber, J. (2006). Canadian listeriosis reference service. *Foodborne Pathogens and Disease*, 3(1):132–137. PMID: 16602988.

Palumbo, J. D., Borucki, M. K., Mandrell, R. E., and Gorski, L. (2003). Serotyping of *Listeria monocytogenes* by enzyme-linked immunosorbent assay and identification of mixed-serotype cultures by colony immunoblotting. *Journal of Clinical Microbiology*, 41(2):564–571.

Paradis, E., Claude, J., and Strimmer, K. (2004). APE: analyses of phylogenetics and evolution in R language. *Bioinformatics*, 20:289–290.

Parsons, M. B., Cooper, K. L. F., Kubota, K. A., Puhr, N., Simington, S., Calimlim, P. S., Schoonmaker-Bopp, D., Bopp, C., Swaminathan, B., Gerner-Smidt, P., and Ribot, E. M. (2007).

PulseNet USA standardized pulsed-field gel electrophoresis protocol for subtyping of *Vibrio parahaemolyticus*. *Foodborne Pathogens and Disease*, 4(3):285–292. PMID: 17883312.

Pearson, T., Okinaka, R. T., Foster, J. T., and Keim, P. (2009). Phylogenetic understanding of clonal populations in an era of whole genome sequencing. *Infection, Genetics and Evolution*, 9(5):1010–1019.

Perna, N. T., Plunkett, G., Burland, V., Mau, B., Glasner, J. D., Rose, D. J., Mayhew, G. F., Evans, P. S., Gregor, J., Kirkpatrick, H. A., Posfai, G., Hackett, J., Klink, S., Boutin, A., Shao, Y., Miller, L., Grotbeck, E. J., Davis, N. W., Lim, A., Dimalanta, E. T., Potamousis, K. D., Apodaca, J., Anantharaman, T. S., Lin, J., Yen, G., Schwartz, D. C., Welch, R. A., and Blattner, F. R. (2001). Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. *Nature*, 409(6819):529–533.

Piffaretti, J. C., Kressebuch, H., Aeschbacher, M., Bille, J., Bannerman, E., Musser, J. M., Selander, R. K., and Rocourt, J. (1989). Genetic characterization of clones of the bacterium *Listeria monocytogenes* causing epidemic disease. *Proceedings of the National Academy of Sciences*, 86(10):3818–3822.

Pinto, F. R., Melo-Cristino, J., and Ramirez, M. (2008). A confidence interval for the wallace coefficient of concordance and its application to microbial typing methods. *PLoS ONE*, 3(11):e3696.

Pop, M. and Salzberg, S. L. (2008). Bioinformatics challenges of new sequencing technology. *Trends in Genetics: TIG*, 24(3):142–149. PMID: 18262676.

R Core Team (2012). R: A language and environment for statistical computing.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850.

Rasko, D. A., Webster, D. R., Sahl, J. W., Bashir, A., Boisen, N., Scheutz, F., Paxinos, E. E., Sebra, R., Chin, C.-S., Iliopoulos, D., Klammer, A., Peluso, P., Lee, L., Kislyuk, A. O., Bullard, J., Kasarskis, A., Wang, S., Eid, J., Rank, D., Redman, J. C., Steyert, S. R., Frimodt-Mller, J., Struve, C., Petersen, A. M., Krogfelt, K. A., Nataro, J. P., Schadt, E. E., and Waldor, M. K. (2011). Origins of the *E. coli* strain causing an outbreak of hemolyticuremic syndrome in Germany. *The New England Journal of Medicine*, 365(8):709–717. PMID: 21793740 PMCID: PMC3168948.

Rasmussen, O. F., Skouboe, P., Dons, L., Rossen, L., and Olsen, J. E. (1995). *Listeria monocytogenes* exists in at least three evolutionary lines: evidence from flagellin, invasive associated protein and listeriolysin O genes. *Microbiology*, 141(9):2053–2061. PMID: 7496516.

Reimer, A. R., Van Domselaar, G., Stroika, S., Walker, M., Kent, H., Tarr, C., Talkington, D., Rowe, L., Olsen-Rasmussen, M., Frace, M., Sammons, S., Dahourou, G. A., Boncy, J., Smith, A. M., Mabon, P., Petkau, A., Graham, M., Gilmour, M. W., and Gerner-Smidt, P. (2011). Comparative genomics of *Vibrio cholerae* from Haiti, Asia, and Africa. *Emerging Infectious Diseases*, 17(11):2113–2121. PMID: 22099115 PMCID: PMC3310578.

Ribot, E. M., Fair, M. A., Gautom, R., Cameron, D. N., Hunter, S. B., Swaminathan, B., and Barrett, T. J. (2006). Standardization of pulsed-field gel electrophoresis protocols for the subtyping of *Escherichia coli* O157:H7, *Salmonella*, and *Shigella* for PulseNet. *Foodborne Pathogens and Disease*, 3(1):59–67. PMID: 16602980.

Robinson, D. F. and Foulds, L. R. (1981). Comparison of phylogenetic trees. *Mathematical Biosciences*, 53(1-2):131–147.

Rohde, H., Qin, J., Cui, Y., Li, D., Loman, N. J., Hentschke, M., Chen, W., Pu, F., Peng, Y., Li, J., Xi, F., Li, S., Li, Y., Zhang, Z., Yang, X., Zhao, M., Wang, P., Guan, Y., Cen, Z., Zhao, X., Christner, M., Kobbe, R., Loos, S., Oh, J., Yang, L., Danchin, A., Gao, G. F., Song, Y., Li, Y., Yang, H., Wang, J., Xu, J., Pallen, M. J., Wang, J., Aepfelbacher, M., and Yang, R. (2011). Open-source genomic analysis of Shiga-toxin-producing *E. coli* O104:H4. *The New England Journal of Medicine*, 365(8):718–724. PMID: 21793736.

Rozen, S. and Skaletsky, H. (2000). Primer3 on the WWW for general users and for biologist programmers. *Methods in Molecular Biology (Clifton, N.J.)*, 132:365–386. PMID: 10547847.

Salcedo, C., Arreaza, L., Alcalá, B., Fuente, L. d. l., and Vázquez, J. A. (2003). Development of a multilocus sequence typing method for analysis of *Listeria monocytogenes* clones. *Journal of Clinical Microbiology*, 41(2):757–762.

Salmonella Subcommittee of the Nomenclature Committee of the International Society for Microbiology (1934). The genus salmonella lignières, 1900. *The Journal of Hygiene*, 34(3):333–350. PMID: 20475239 PMCID: PMC2170865.

Sanger, F., Nicklen, S., and Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, 74(12):5463 –5467.

Scaria, J., Palaniappan, R. U., Chiu, D., Phan, J. A., Ponnala, L., McDonough, P., Grohn, Y. T., Porwollik, S., McClelland, M., Chiou, C.-S., Chu, C., and Chang, Y.-F. (2008). Microarray for molecular typing of *Salmonella enterica* serovars. *Molecular and Cellular Probes*, 22(4):238–243.

Schena, M., Shalon, D., Davis, R. W., and Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270(5235):467–470. PMID: 7569999.

Schlech, W. F., Lavigne, P. M., Bortolussi, R. A., Allen, A. C., Haldane, E. V., Wort, A. J., Hightower, A. W., Johnson, S. E., King, S. H., Nicholls, E. S., and Broome, C. V. (1983). Epidemic listeriosis evidence for transmission by food. *New England Journal of Medicine*, 308(4):203–206.

Schürch, A. C. and van Soolingen, D. (2012). DNA fingerprinting of *Mycobacterium tuberculosis*: From phage typing to whole-genome sequencing. *Infection, Genetics and Evolution*, 12(4):602–609.

Seeliger, H. and Langer, B. (1989). Serological analysis of the genus *Listeria*. its values and limitations. *International Journal of Food Microbiology*, 8(3):245–248.

Selander, R. K., Caugant, D. A., Ochman, H., Musser, J. M., Gilmour, M. N., and Whittam, T. S. (1986). Methods of multilocus enzyme electrophoresis for bacterial population genetics and systematics. *Applied and Environmental Microbiology*, 51(5):873–884. PMID: 2425735 PMCID: PMC238981.

Selander, R. K., Musser, J. M., Caugant, D. A., Gilmour, M. N., and Whittam, T. S. (1987). Population genetics of pathogenic bacteria. *Microbial Pathogenesis*, 3(1):1–7.

Severiano, A., Carri co, J. a. A., Robinson, D. A., Ramirez, M., and Pinto, F. R. (2011a). Evaluation of jackknife and bootstrap for defining confidence intervals for pairwise agreement measures. *PLoS ONE*, 6(5):e19539.

Severiano, A., Pinto, F. R., Ramirez, M., and Carriço, J. a. A. (2011b). Adjusted Wallace as a measure of congruence between typing methods. *Journal of Clinical Microbiology*, 49(11):3997–4000.

Shendure, J. and Ji, H. (2008). Next-generation DNA sequencing. *Nat Biotech*, 26(10):1135–1145.

Sheneman, L., Evans, J., and Foster, J. A. (2006). Clearcut: a fast implementation of relaxed neighbor joining. *Bioinformatics*, 22(22):2823–2824.

Silbert, S., Boyken, L., Hollis, R. J., and Pfaller, M. A. (2003). Improving typeability of multiple bacterial species using pulsed-field gel electrophoresis and thiourea. *Diagnostic Microbiology and Infectious Disease*, 47(4):619–621. PMID: 14711485.

Simpson, E. (1949). Measurement of diversity. *Nature*, 163(4148):688.

Smith, J. M., Dowson, C. G., and Spratt, B. G. (1991). Localized sex in bacteria. 349(6304):29–31.

Smith, J. M., Smith, N. H., O'Rourke, M., and Spratt, B. G. (1993). How clonal are bacteria? *Proceedings of the National Academy of Sciences*, 90(10):4384–4388.

Snipen, L., Almy, T., and Ussery, D. W. (2009). Microbial comparative pan-genomics using binomial mixture models. *BMC Genomics*, 10:385. PMID: 19691844.

Snow, J. (1857). Cholera, and the water supply in the south districts of London. *British Medical Journal*, 1(42):864–865. PMID: null PMCID: PMC2250686.

Sperry, K. E. V., Kathariou, S., Edwards, J. S., and Wolf, L. A. (2008). Multiple-locus variable-number tandem-repeat analysis as a tool for subtyping *Listeria monocytogenes* strains. *Journal of Clinical Microbiology*, 46(4):1435–1450. PMID: 18256218.

Sreevatsan, S., Pan, X., Stockbauer, K. E., Connell, N. D., Kreiswirth, B. N., Whittam, T. S., and Musser, J. M. (1997). Restricted structural gene polymorphism in the *Mycobacterium tuberculosis* complex indicates evolutionarily recent global dissemination. *Proceedings of the National Academy of Sciences of the United States of America*, 94(18):9869–9874. PMID: 9275218.

Stajich, J. E., Block, D., Boulez, K., Brenner, S. E., Chervitz, S. A., Dagdigian, C., Fuellen, G., Gilbert, J. G., Korf, I., Lapp, H., Lehväslaiho, H., Matsalla, C., Mungall, C. J., Osborne, B. I., Pocock, M. R., Schattner, P., Senger, M., Stein, L. D., Stupka, E., Wilkinson, M. D., and

Birney, E. (2002). The BioPerl toolkit: Perl modules for the life sciences. *Genome Research*, 12(10):1611 –1618.

Sugimoto, N., Shima, K., Hinenoya, A., Asakura, M., Matsuhisa, A., Watanabe, H., and Yamasaki, S. (2011). Evaluation of a PCR-restriction fragment length polymorphism (PCR-RFLP) assay for molecular epidemiological study of Shiga toxin-producing *Escherichia coli*. *The Journal of Veterinary Medical Science / the Japanese Society of Veterinary Science*, 73(7):859–867. PMID: 21321474.

Sul, S.-J., Brammer, G., and Williams, T. L. (2008). Efficiently computing arbitrarily-sized Robinson-Foulds distance matrices. In *Proceedings of the 8th international workshop on Algorithms in Bioinformatics*, WABI '08, pages 123–134, Berlin, Heidelberg. Springer-Verlag.

Swaminathan, B., Barrett, T. J., Hunter, S. B., and Tauxe, R. V. (2001). PulseNet: the molecular subtyping network for foodborne bacterial disease surveillance, United States. *Emerging Infectious Diseases*, 7(3):382–389. PMID: 11384513.

Swaminathan, B. and Gerner-Smidt, P. (2007). The epidemiology of human listeriosis. *Microbes and Infection*, 9(10):1236–1243.

Taboada, E. N., Acedillo, R. R., Carrillo, C. D., Findlay, W. A., Medeiros, D. T., Mykytczuk, O. L., Roberts, M. J., Valencia, C. A., Farber, J. M., and Nash, J. H. E. (2004). Large-scale comparative genomics meta-analysis of *Campylobacter jejuni* isolates reveals low level of genome plasticity. *Journal of Clinical Microbiology*, 42(10):4566–4576. PMID: 15472310.

Taboada, E. N., Luebbert, C. C., and Nash, J. H. E. (2007). Studying bacterial genome dynamics using microarray-based comparative genomic hybridization. *Methods in Molecular Biology (Clifton, N.J.)*, 396:223–253. PMID: 18025696.

Taboada, E. N., Mackinnon, J. M., Luebbert, C. C., Gannon, V. P. J., Nash, J. H. E., and Rahn, K. (2008). Comparative genomic assessment of multi-locus sequence typing: rapid accumulation of genomic heterogeneity among clonal isolates of *Campylobacter jejuni*. *BMC Evolutionary Biology*, 8:229. PMID: 18691421.

Taboada, E. N., Ross, S. L., Mutschall, S. K., Mackinnon, J. M., Roberts, M. J., Buchanan, C. J., Kruczkiewicz, P., Jokinen, C. C., Thomas, J. E., Nash, J. H. E., Gannon, V. P. J., Marshall, B., Pollari, F., and Clark, C. G. (2012). Development and validation of a comparative

genomic fingerprinting method for high-resolution genotyping of *Campylobacter jejuni*. *Journal of Clinical Microbiology*, 50(3):788–797. PMID: 22170908.

Tettelin, H., Masignani, V., Cieslewicz, M. J., Donati, C., Medini, D., Ward, N. L., Angiuoli, S. V., Crabtree, J., Jones, A. L., Durkin, A. S., Deboy, R. T., Davidsen, T. M., Mora, M., Scarselli, M., Margarit y Ros, I., Peterson, J. D., Hauser, C. R., Sundaram, J. P., Nelson, W. C., Madupu, R., Brinkac, L. M., Dodson, R. J., Rosovitz, M. J., Sullivan, S. A., Daugherty, S. C., Haft, D. H., Selengut, J., Gwinn, M. L., Zhou, L., Zafar, N., Khouri, H., Radune, D., Dimitrov, G., Watkins, K., O'Connor, K. J. B., Smith, S., Utterback, T. R., White, O., Rubens, C. E., Grandi, G., Madoff, L. C., Kasper, D. L., Telford, J. L., Wessels, M. R., Rappuoli, R., and Fraser, C. M. (2005). Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". *Proceedings of the National Academy of Sciences of the United States of America*, 102(39):13950–13955. PMID: 16172379.

Tettelin, H., Riley, D., Cattuto, C., and Medini, D. (2008). Comparative genomics: the bacterial pan-genome. *Current Opinion in Microbiology*, 11(5):472–477.

Van Belkum, A., Tassios, P. T., Dijkshoorn, L., Haeggman, S., Cookson, B., Fry, N. K., Fussing, V., Green, J., Feil, E., Gerner-Smidt, P., Brisse, S., Struelens, M., Microbiology, f. t. E. S. o. C., and (esgem), I. D. E. S. G. o. E. M. (2007). Guidelines for the validation and application of typing methods for use in bacterial epidemiology. *Clinical Microbiology and Infection*, 13:1–46.

Van Stelten, A. and Nightingale, K. K. (2008). Development and implementation of a multiplex single-nucleotide polymorphism genotyping assay for detection of virulence-attenuating mutations in the *Listeria monocytogenes* virulence-associated gene inlA. *Applied and Environmental Microbiology*, 74(23):7365–7375. PMID: 18836010.

Van Stelten, A., Simpson, J. M., Ward, T. J., and Nightingale, K. K. (2010). Revelation by single-nucleotide polymorphism genotyping that mutations leading to a premature stop codon in inlA are common among *Listeria monocytogenes* isolates from ready-to-eat foods but not human listeriosis cases. *Applied and Environmental Microbiology*, 76(9):2783–2790.

Vaneechoutte, M., Boerlin, P., Tichy, H.-V., Bannerman, E., Jger, B., and Bille, J. (1998). Comparison of PCR-based DNA fingerprinting techniques for the identification of *Listeria* species and their use for atypical *Listeria* isolates. *International Journal of Systematic Bacteriology*, 48(1):127–139.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., Skupski, M., Subramanian, G., Thomas, P. D., Zhang, J., Gabor Miklos, G. L., Nelson, C., Broder, S., Clark, A. G., Nadeau, J., McKusick, V. A., Zinder, N., Levine, A. J., Roberts, R. J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Di Francesco, V., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A. E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T. J., Higgins, M. E., Ji, R. R., Ke, Z., Ketchum, K. A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G. V., Milshina, N., Moore, H. M., Naik, A. K., Narayan, V. A., Neelam, B., Nusskern, D., Rusch, D. B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z., Wang, A., Wang, X., Wang, J., Wei, M., Wides, R., Xiao, C., Yan, C., Yao, A., Ye, J., Zhan, M., Zhang, W., Zhang, H., Zhao, Q., Zheng, L., Zhong, F., Zhong, W., Zhu, S., Zhao, S., Gilbert, D., Baumhueter, S., Spier, G., Carter, C., Cravchik, A., Woodage, T., Ali, F., An, H., Awe, A., Baldwin, D., Baden, H., Barnstead, M., Barrow, I., Beeson, K., Busam, D., Carver, A., Center, A., Cheng, M. L., Curry, L., Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferriera, S., Garg, N., Gluecksmann, A., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L., Koduru, S., Love, A., Mann, F., May, D., McCawley, S., McIntosh, T., McMullen, I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratts, E., Puri, V., Qureshi, H., Reardon, M., Rodriguez, R., Rogers, Y. H., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N. N., Tse, S., Vech, C., Wang, G., Wetter, J., Williams, S., Williams, M., Windsor, S., Winn-Deen, E., Wolfe, K., Zaveri, J., Zaveri, K., Abril, J. F., Guig, R., Campbell, M. J., Sjolander, K. V., Karlak, B., Kejariwal, A., Mi, H., Lazareva, B., Hatton, T., Narechania, A., Diemer, K., Muruganujan, A., Guo, N., Sato, S., Bafna, V., Istrail, S., Lippert, R., Schwartz, R., Walenz, B., Yooseph, S., Allen, D., Basu, A., Baxendale, J., Blick, L., Caminha, M., Carnes-Stine, J., Caulk, P., Chiang, Y. H., Coyne, M., Dahlke, C., Mays, A., Dombroski, M., Donnelly, M., Ely, D., Esparham, S., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, A., Gorokhov, M., Graham, K., Gropman, B., Harris, M., Heil, J., Henderson, S., Hoover, J., Jennings, D., Jordan, C., Jordan,

J., Kasha, J., Kagan, L., Kraft, C., Levitsky, A., Lewis, M., Liu, X., Lopez, J., Ma, D., Majoros, W., McDaniel, J., Murphy, S., Newman, M., Nguyen, T., Nguyen, N., Nodell, M., Pan, S., Peck, J., Peterson, M., Rowe, W., Sanders, R., Scott, J., Simpson, M., Smith, T., Sprague, A., Stockwell, T., Turner, R., Venter, E., Wang, M., Wen, M., Wu, D., Wu, M., Xia, A., Zandieh, A., and Zhu, X. (2001). The sequence of the human genome. *Science*, 291(5507):1304–1351. PMID: 11181995.

Vergnaud, G. and Pourcel, C. (2009). Multiple locus variable number of tandem repeats analysis. *Methods in Molecular Biology*, 551:141–158. PMID: 19521873.

Versalovic, J., Kapur, V., Mason, E O, J., Shah, U., Koeuth, T., Lupski, J. R., and Musser, J. M. (1993). Penicillin-resistant *Streptococcus pneumoniae* strains recovered in Houston: identification and molecular characterization of multiple clones. *The Journal of Infectious Diseases*, 167(4):850–856. PMID: 8450250.

Vinten-Johansen, P. (2003). *Cholera, Chloroform, and the Science of Medicine: A Life of John Snow*. Oxford University Press.

Vos, P., Hogers, R., Bleeker, M., Reijans, M., van de Lee, T., Hornes, M., Frijters, A., Pot, J., Peleman, J., and Kuiper, M. (1995). AFLP: a new technique for DNA fingerprinting. *Nucleic Acids Research*, 23(21):4407–4414. PMID: 7501463 PMCID: PMC307397.

Wallace, D. L. (1983). A method for comparing two hierarchical clusterings: Comment. *Journal of the American Statistical Association*, 78(383):569.

Wang, G., Whittam, T. S., Berg, C. M., and Berg, D. E. (1993). RAPD (arbitrary primer) PCR is more sensitive than multilocus enzyme electrophoresis for distinguishing related bacterial strains. *Nucleic Acids Research*, 21(25):5930–5933.

Ward, T. J., Ducey, T. F., Usgaard, T., Dunn, K. A., and Bielawski, J. P. (2008). Multilocus genotyping assays for single nucleotide polymorphism-based subtyping of *Listeria monocytogenes* isolates. *Applied and Environmental Microbiology*, 74(24):7629–7642. PMID: 18931295.

Ward, T. J., Usgaard, T., and Evans, P. (2010). A targeted multilocus genotyping assay for lineage, serogroup, and epidemic clone typing of *Listeria monocytogenes*. *Applied Environmental Microbiology*, 76(19):6680–6684.

Wentworth, B. B. (1963). Bacteriophage typing of the staphylococci. *Bacteriological Reviews*, 27(3):253–272. PMID: 14063853 PMCID: PMC441186.

Whittam, T. S., Ochman, H., and Selander, R. K. (1983). Geographic components of linkage disequilibrium in natural populations of *Escherichia coli*. *Molecular Biology and Evolution*, 1(1):67–83. PMID: 6400648.

Wiedmann, M., Bruce, J. L., Keating, C., Johnson, A. E., McDonough, P. L., and Batt, C. A. (1997). Ribotypes and virulence gene polymorphisms suggest three distinct *Listeria monocytogenes* lineages with differences in pathogenic potential. *Infection and Immunity*, 65(7):2707–2716. PMID: 9199440 PMCID: 175382.

Williams, J. G., Kubelik, A. R., Livak, K. J., Rafalski, J. A., and Tingey, S. V. (1990). DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. *Nucleic Acids Research*, 18(22):6531–6535. PMID: 1979162.

Xiao, X., Yan, Y., Zhang, Y., Wang, L., Liu, X., Yang, L., Tan, Y., Guo, Z., Yang, R., and Zhou, D. (2011). A novel genotyping scheme for *Vibrio parahaemolyticus* with combined use of large variably-presented gene clusters (LVPCs) and variable-number tandem repeats (VNTRs). *International Journal of Food Microbiology*, 149(2):143–151.

Xu, Z., Chen, X., Li, L., Li, T., Wang, S., Chen, H., and Zhou, R. (2010). Comparative genomic characterization of *Actinobacillus pleuropneumoniae*. *Journal of Bacteriology*. PMID: 20802045.

Yde, M. and Genicot, A. (2004). Use of PFGE to characterize clonal relationships among Belgian clinical isolates of *Listeria monocytogenes*. *Journal of Medical Microbiology*, 53(Pt 5):399–402. PMID: 15096548.

Yoshida, C., Franklin, K., Konczy, P., McQuiston, J. R., Fields, P. I., Nash, J. H., Taboada, E. N., and Rahn, K. (2007). Methodologies towards the development of an oligonucleotide microarray for determination of *Salmonella* serotypes. *Journal of Microbiological Methods*, 70(2):261–271.

Zhang, C., Zhang, M., Ju, J., Nietfeldt, J., Wise, J., Terry, P. M., Olson, M., Kachman, S. D., Wiedmann, M., Samadpour, M., and Benson, A. K. (2003). Genome diversification in phylogenetic lineages I and II of *Listeria monocytogenes*: identification of segments unique to lineage II populations. *Journal of Bacteriology*, 185(18):5573–5584. PMID: 12949110.

Zhang, W., Jayarao, B. M., and Knabel, S. J. (2004). Multi-virulence-locus sequence typing of *Listeria monocytogenes*. *Applied and Environmental Microbiology*, 70(2):913–920. PMID: 14766571 PMCID: 348834.

Zhang, Y., Laing, C., Steele, M., Ziebell, K., Johnson, R., Benson, A. K., Taboada, E., and Gannon, V. P. (2007). Genome evolution in major *Escherichia coli* O157:H7 lineages. *BMC Genomics*, 8(1):121.

Listeria monocytogenes strain information

TABLE 1: The 37 Canadian outbreak-related *L. monocytogenes* strains included in the design and validation of a *L. monocytogenes* CGF assay are shown here.

Strain	Lineage	Serotype	Year	Province	Source	Sample Source
10-4754	II	1/2a	2002	Quebec	Cheese	Cerebrospinal Fluid
10-4758	II	1/2a	2002	Quebec	Cheese	Cheese
10-0812	II	1/2a	2000	Manitoba	Whipping Cream	Whipping cream
10-0813	II	1/2a	2000	Manitoba	Whipping Cream	Clinical
10-0933	II	1/2a	2008	Quebec/Ontario	Cheese	Blood
10-0934	II	1/2a	2008	Quebec/Ontario	Cheese	Cheese
10-5025	II	1/2c	2010		Food Industry	Bacon Flakes
10-5026	II	1/2c	2010		Food Industry	Bacon Flake Processing Plant
10-5027	II	3c	2010		Food Industry	Processing Plant
10-1046	II	1/2a	2010	Ontario	Prosciutto Ham	Blood
10-1047	II	1/2a	2010	Ontario	Prosciutto Ham	Cerebrospinal Fluid
10-1321	II	1/2a	2010	Ontario	Prosciutto Ham	Blood
10-5024	II	1/2a	2010	Ontario	Prosciutto Ham	Prosciutto Ham
02-5993	II	1/2a	2002		ECV Isolates	Blood
88-0478	II	1/2a	1988		ECV Isolates	Blood
08-7669	II	1/2a	2008		Ready-To-Eat Meat	Blood
95-0093	II	1/2a	1995		ECV Isolates	Blood
98-2035	II	1/2a	1998		LGI1 positive	Blood
99-6370	II	1/2a	1999		LGI1 positive	Blood
04-5457	II	1/2a	2004		ECV Isolates	Blood
08-6056	II	1/2a	2008		Ready-To-Eat Meat	Turkey Meat
08-6569	II	1/2a	2008		Ready-To-Eat Meat	Environmental
08-6997	II	1/2a	2008		Ready-To-Eat Meat	Blood
08-7374	II	1/2a	2008		Ready-To-Eat Meat	Environmental
10-0814	II	1/2a	2008		Ready-To-Eat Meat	Ready-To-Eat Meat
10-0815	II	1/2a	2008		Ready-To-Eat Meat	Ready-To-Eat Meat
10-0810	I	1/2b	1996	Ontario	Imitation Crab	Clinical
10-0811	I	1/2b	1996	Ontario	Imitation Crab	Imitation Crab
02-6679	I	4b	2002	British Columbia	Cheese	Stool
02-6680	I	4b	2002	British Columbia	Cheese	Cheese
81-0558	I	4b	1981	Maritimes	Coleslaw	Cerebrospinal Fluid
10-0809	I	4b	1981	Maritimes	Coleslaw	Clinical
81-0592	I	4b	1981	Maritimes	Coleslaw	Fetal Blood
81-0861	I	4b	1981	Maritimes	Coleslaw	Coleslaw
02-1103	I	4b	2002	British Columbia	Cheese	Cerebrospinal Fluid
02-1289	I	4b	2002	British Columbia	Cheese	Stool
02-1792	I	4b	2002	British Columbia	Cheese	Cheese

TABLE 2: The 23 publicly available *L. monocytogenes* strains retrieved from NCBI included in the design and validation of a *L. monocytogenes* CGF assay are shown here.

Strain	Lineage	Serotype	Taxonomy ID
J1-194	I	1/2b	393117
N1-017	I	4b	393123
R2-503	I	1/2b	393125
Clip80459	I	4b	568819
H7858	I	4b	267410
F2365	I	4b	265669
HPB2262	I	4b	401650
Scott-A	I	4b	1027396
F6854	II	1/2a	267409
F6900	II	1/2a	393128
J0161	II	1/2a	393130
J2818	II	1/2a	393131
N3-165	II	1/2a	393124
1988	II	3a	393127
10403S	II	1/2a	393133
EGD-e	II	1/2a	169963
R2-561	II	1/2c	393126
08-5923	II	1/2a	637381
08-5578	II	1/2a	653938
HCC23	III	4a	552536
L99-4a	III	4a	563174
M7	III	4a	1030009
J2-071	IIIA	4c	393121

Thesis software project source code repositories

Table 3: Links to the source code repositories for the software applications used in this thesis.

Application Name	Abbreviation	URL Link
CGF Optimizer	CGFO	`https://bitbucket.org/peterk87/cgfoptimizer`
CGF Multiplexer	CGFM	`https://bitbucket.org/peterk87/cgfmultiplexer`
Microbial *In Silico* Typer	MIST	`https://bitbucket.org/peterk87/microbialinsilicotyper`

L. monocytogenes CGF40 marker information

TABLE 4: *L. monocytogenes* CGF marker protein products are shown here. Protein products for each marker were determined using NCBI BLAST. Dashes denote hypothetical proteins.

CGF Marker ID	Protein Product
02_1103-0112	-
02_1103-0160	*Not found in L. monocytogenes*
02_1103-0189	ROK family protein
02_1103-0355	Internalin
02_1103-0525	-
02_1103-0540	-
02_1103-0717	PTS system IIA 2 domain-containing protein
02_1103-0991	-
02_1103-1213	Regulatory protein
02_1103-2432	-
02_1103-2509	Carboxylesterase
02_5993-0412	-
02_5993-0547	Type I restriction enzyme, R subunit
02_5993-0787	Conserved hypothetical protein
02_5993-0853	Protease synthase and sporulation negative regulatory protein PAI 1
02_5993-1121	-
02_5993-1275	-
02_5993-1292	-
02_5993-1325	-
02_5993-1753	-
02_5993-2804	Internalin protein
02_6679-0315	HsdR family type I site-specific deoxyribonuclease
02_6679-0490	-
10403S-20100007802	-
10403S-20100013339	-
10_0812-1247	Phage protein
10_0933-0083	-
10_4754-1779	Prophage Lp2 protein 6
10_5025-1274	*Not found in L. monocytogenes*
Clip80459-0475	-
F6854-2467	DEAD/DEAH box helicase domain-containing protein
Fin1988-10100012229	ATP-dependent RNA helicase
H7858-0863	Oxidoreductase
H7858-2421	Phage minor structural protein, N-terminal region subfamily
J0161-20100013416	NADH oxidase
J0161-20100014638	-
J0161-20100015457	IS30 family, transposase
J1_194-2981	Conserved hypothetical protein
N1_017-LMHG	gp36 protein
R2_561-10100005007	gp52 protein

The CGF markers, J0161-20100013416 and J0161-20100015457, were found on plasmids. J0161-20100013416 was found on the *L. monocytogenes* 08-5578 plasmid pLM5578 and on the *Listeria innocua* Clip11262 plasmid pLI100. J0161-20100015457 was found on the *L. monocytogenes* H7858 plasmid pLM80. Since these markers were found on plasmids, a DNA extraction protocol capable of reliably extracting genomic DNA and plasmid DNA simultaneously may be necessary.

Two of the CGF markers, 02_1103-0160 and 10_5025-1274, could not be found in *L. monocytogenes* using NCBI BLAST. The closest match to 02_1103-0160 was within *Listeria grayi* DSM 20601 as a GNAT family acetyltransferase. The closest match to 10_5025-1274 was within *Streptococcus suis* GZ1 as a hypothetical protein with only 33% ID.

TABLE 5: *L. monocytogenes* CGF markers and PCR primers are shown here.

ORF	Multiplex	CGF Marker ID	Forward	Reverse	Product Size (bp)
LMOf6854_2467-00233489-F6854	1	F6854-2467	CCGCAAATGTATCTGATTTTGA	TTTTCTGCACCAAACAAATGAC	154
0853-0853-02_5993	1	02_5993-0853	TGATTCCCAAGCACTCTTAGAA	GCTCTATTTCGATTGTATTCCCA	251
0189-0189-02_1103	1	02_1103-0189	TAGATACGGAAAAAGGCGTGAT	GCTTAATTAACTCCACCGCTGA	351
2432-2432-02_1103	1	02_1103-2432	AGTTAAAAGAAATTGCCGACGA	TTTCTTTTGAGCTCTTTTTGCAT	470
1121-1121-02_5993	1	02_5993-1121	ACCAGGCATTAGCCATACAGAT	TGTTGTCCCGTTTAAAATACCC	569
1325-1325-02_5993	2	02_5993-1325	GGAGAGGTCCTGTAATGAAACG	TCTCCCAATCTCTAGTCCCAAA	212
Lmon1_020100007802-05235900-10403S	2	10403S-20100007802	TCGCAAGAGAAGAAAGAACAAA	CCTCTGGTAGTGACTTGTGCAT	323
LmonFR_010100005007-03670169-R2_561	2	R2_561-10100005007	AAAGAATGGTAGAAATCAACGGA	TTCGCGAAATACTTTCCATAAAA	409
0525-0525-02_1103	2	02_1103-0525	TGATTATGATGAAGAAAAGAATCCA	TTTAAATTTCCTAGGTGTGCGG	558
2509-2509-02_1103	2	02_1103-2509	TAGCTTTACGTTAAAAGCGGGA	AGCCAAACTAAATCGTCTAAGAAGT	711
1753-1753-02_5993	3	02_5993-1753	GCATTAGAGCGGTCTCATACATA	ATTTCAGGATTATTTCCCCACA	169
0355-0355-02_1103	3	02_1103-0355	CATCAGTGTCCTTTCAGAATGC	TTTTGTCCAGTAGCATCAAACAA	252
0412-0412-02_5993	3	02_5993-0412	TTGAATCAAAAATAAAGTCTCAAGGA	CCATTTTCATCAATTCCCATTA	403
0160-0160-02_1103	3	02_1103-0160	TCCTCTCTTCTTTTGAGTGCAA	TGGTGACTCGTAATCTTCTAAAAGTC	500
LMBG_02981-05231399-J1_194	3	J1_194-2981	ATCGAAAAATATGGAAGAAGCG	AACCAGCTTCTAAACTCCCGTA	620
0540-0540-02_1103	4	02_1103-0540	TGAATTTGATTTATTAACAGAGGATGA	GCAAAAGTGTTTGGGCATACTT	222
0717-0717-02_1103	4	02_1103-0717	TGAAAACGCTCTCTTGAATGAA	ATTGGAAGTACCAGCTTTGCAT	300
0547-0547-02_5993	4	02_5993-0547	ACGAGGCCAAAATGATCTAAAA	GTTTCCTGAAGAGAATCGGTTG	474
0991-0991-02_1103	4	02_1103-0991	GAAATTGCCCTTTGATTTGTC	TTGGCTCACTTCCTGTTAATTC	575
1274-1274-10_5025	4	10_5025-1274	TGGAGATGACATGGTTAAAGAAA	GCTCGCAAATCTCCTTCAACT	723
0490-0490-02_6679	5	02_6679-0490	TTCCATAATCCGGAAATTAATGAT	AGACCCGTTTAAAGGTGACTGA	210
1275-1275-02_5993	5	02_5993-1275	AGCAGTTGGTGGTTTTAATGCT	CTCCTTTTTCATCGTTTCCATC	362
0787-0787-02_5993	5	02_5993-0787	AATGAAGCAAAATCAATACGGC	TCTTCAGCATTTTCTACTTCAATTTT	462
0315-0315-02_6679	5	02_6679-0315	CATATGCCGAGAGTTGTTGAAG	TCGTCCTCTGTCATTCTAGGGT	554
LmonJ_020100014638-05260984-J0161	5	J0161-20100014638	GAAATGGAGGAAGTAAAAGGGC	TTCAGATTTCGAAGAGTCCGTT	720
0083-0083-10_0933	6	10_0933-0083	TGTTGCATACTAGAGTTGTTGGTG	CACATCCACATAAGGCACATTT	152
1213-1213-02_1103	6	02_1103-1213	AAGTGGAAAAGGCGATTATTGA	ACCCTTCTCAAATTCATCAGGT	275
1292-1292-02_5993	6	02_5993-1292	ATCTTTAACAACGAGCCGGTAA	TGCGATTGAAAGTTGATATAGGG	400
LMOh7858_2421-00230175-H7858	6	H7858-2421	CTTCTGAACTTAAGGGAAGCCA	ACATATCCTGCATTTCCTAGCC	571
2804-2804-02_5993	6	02_5993-2804	AGCGACGATAATAGCAAAAAGC	TGTGCTATCTCCTGTTTCTGGA	715
1247-1247-10_0812	7	10_0812-1247	TTCTAATAGGCGGAAATGTCGT	GCGTCTCTCTACCGAACCTATC	174
LMHG_11509-07075031-N1_017	7	N1_017-LMHG	AAAGACGTGGCAAAAGTATTGG	TGATTGCTTTTTCGATTGTGTC	271
Lmon1_020100013339-05236987-10403S	7	10403S-20100013339	TTGACGAATTAAATAAAATATTATCGG	TTATTTTGTTGTAAAAACATCTCCAAA	400
0112-0112-02_1103	7	02_1103-0112	AAAGGATGCTTTAATCGCAGAA	CATTCACTTTTCCAACCATCAA	554
LmonJ_020100013416-05260741-J0161	7	J0161-20100013416	ACAATCCCGAATGTAGATGACC	TTGGTAAGAGTTTGCGTTAGCA	705
Lm4b_00457-002757168-Clip80459	8	Clip80459-0475	CAATTTGTTTCATTCCGCTG	TTTGCACTACCAAAATCACATAAA	165
1779-1779-10_4754	8	10_4754-1779	CGCCATTCTTTTTCCTTTGTACT	AAATTGTCGAAATCCCCTTTTT	275
LmonJ_020100015457-05261139-J0161	8	J0161-20100015457	GCTTTTGACGTTTCCCTTTCT	GAACGCGTGAAAATAGAGACCT	361
LMOh7858_0863-00231670-H7858	8	H7858-0863	CCCGAAATTGAAGTCCTAATCA	TGGGCCATTTCATCACTTGTAT	500
LmonF1_010100012229-03668733-1988	8	Fin1988-10100012229	AAATTCCGCATGGTTATTTGTC	TTGTGATTCGCTCATAAATTGC	621

In silico typing data

TABLE 6: Cluster numbers defined at fingerprint similarities of 100, 97.5, 95 and 90% are shown here. CGF40 fingerprints were derived from *in silico* typing of the 60 *L. monocytogenes* strains.

Strains	100%	97.5%	95%	90%
EGD-e	1	1	1	1
F2365	2	2	2	2
F6854	3	3	3	3
F6900	4	4	4	3
H7858	5	5	5	4
HCC23	6	6	6	5
HPB2262	7	7	7	6
J1-194	8	8	8	7
J2-071	9	9	9	8
J0161	10	10	10	9
J2818	11	4	4	3
L99-4a	12	11	11	5
M7	12	11	11	5
N1-017	13	12	12	10
N3-165	14	13	13	11
R2-503	15	14	14	12
R2-561	16	15	15	13
Scott-A	17	16	16	14
02-1103	18	17	17	2
02-1289	18	17	17	2
02-1792	18	17	17	2
02-5993	19	18	18	15
04-5457	19	18	18	15
08-5578	19	18	18	15
08-5923	19	18	18	15
08-6056	19	18	18	15
08-6569	19	18	18	15

Continued on next page

Strains	100%	97.5%	95%	90%
08-6997	19	18	18	15
08-7374	19	18	18	15
10-0814	19	18	18	15
10-0815	19	18	18	15
10-1321	19	18	18	15
10-5024	19	18	18	15
98-2035	19	18	18	15
99-6370	19	18	18	15
02-6679	20	19	19	14
02-6680	20	19	19	14
08-7669	21	20	20	15
10-0809	22	21	2	2
81-0558	22	21	2	2
81-0592	22	21	2	2
81-0861	22	21	2	2
10-0810	23	22	21	16
10-0811	23	22	21	16
10-0812	24	23	22	17
10-0813	24	23	22	17
10-0933	25	24	23	18
10-0934	25	24	23	18
10-1046	26	25	24	15
10-1047	26	25	24	15
10-4754	27	26	25	19
10-4758	28	26	25	19
10-5025	29	27	26	13
10-5026	29	27	26	13
10-5027	30	28	15	13
88-0478	31	25	24	15
95-0093	32	18	18	15
1988	33	29	27	20
10403S	34	30	28	21

Continued on next page

Strains	100%	97.5%	95%	90%
Clip80459	35	31	29	22

TABLE 7: The *in silico*-derived MLST data for 60 *L. monocytogenes* strains are shown here. *In silico*-derived allele designations for each MLST gene locus are shown for each strain and the resulting sequence type (ST). The allele designation for the *dat* MLST locus could not be determined for H7858 due to a truncation of the locus by the end of a contig.

Strain	*abcZ*	*bglA*	*cat*	*dapE*	*dat*	*ldh*	*lhkA*	ST
02-1103	3	1	1	1	3	1	3	1
02-1289	3	1	1	1	3	1	3	1
02-1792	3	1	1	1	3	1	3	1
02-5993	5	6	2	29	5	3	1	120
02-6679	3	1	12	70	2	1	5	388
02-6680	3	1	12	70	2	1	5	388
04-5457	57	6	2	29	5	3	1	292
08-5578	57	6	2	29	5	3	1	292
08-5923	5	6	2	29	5	3	1	120
08-6056	57	6	2	29	5	3	1	292
08-6569	57	6	2	29	5	3	1	292
08-6997	57	6	2	29	5	3	1	292
08-7374	57	6	2	29	5	3	1	292
08-7669	5	6	2	29	5	3	1	120
10403S	5	8	5	7	6	38	1	85
10-0809	3	1	1	1	3	1	3	1
10-0810	2	1	11	3	3	1	7	5
10-0811	2	1	11	3	3	1	7	5
10-0812	5	8	5	7	6	2	1	7
10-0813	5	8	5	7	6	2	1	7
10-0814	57	6	2	29	5	3	1	292
10-0815	57	6	2	29	5	3	1	292
10-0933	5	5	17	21	39	2	6	394
10-0934	5	5	17	21	39	2	6	394
10-1046	5	6	2	29	5	3	1	120
10-1047	5	6	2	29	5	3	1	120
10-1321	5	6	2	29	5	3	1	120
10-4754	5	7	3	5	1	8	6	37
10-4758	5	7	3	5	1	8	6	37

Continued on next page

Strain	*abcZ*	*bglA*	*cat*	*dapE*	*dat*	*ldh*	*lhkA*	ST
10-5024	5	6	2	29	5	3	1	120
10-5025	6	5	6	4	1	4	1	9
10-5026	6	5	6	4	1	4	1	9
10-5027	6	5	6	4	1	4	1	9
1988	7	10	16	7	5	2	1	155
81-0558	3	1	1	1	3	1	3	1
81-0592	3	1	1	1	3	1	3	1
81-0861	3	1	1	1	3	1	3	1
88-0478	5	6	2	29	5	3	1	120
95-0093	5	6	2	29	5	3	1	120
98-2035	57	6	2	29	5	3	1	292
99-6370	57	6	2	29	5	3	1	292
Clip80459	1	2	12	3	2	5	3	4
EGD-e	6	5	6	20	1	4	1	35
F2365	3	1	1	1	3	1	3	1
F6854	7	6	10	6	1	2	1	11
F6900	7	6	10	6	1	2	1	11
H7858	3	9	9	3	-	1	5	-
HCC23	19	17	22	25	15	79	12	201
HPB2262	1	1	11	11	2	1	5	2
J0161	7	6	10	6	1	2	1	11
J1-194	12	12	12	1	3	1	4	88
J2818	7	6	10	6	1	2	1	11
J2-071	18	11	21	24	17	31	13	131
L99-4a	19	17	22	25	15	79	12	201
M7	19	17	22	25	15	79	12	201
N1-017	4	4	4	3	2	1	5	3
N3-165	21	6	15	8	6	2	14	222
R2-503	4	4	4	3	2	1	5	3
R2-561	6	5	6	4	1	4	1	9
Scott-A	1	1	11	11	2	11	5	290

TABLE 8: The *in silico*-derived MVLST data for 60 *L. monocytogenes* strains are shown here. *In silico*-derived allele designations for each MVLST gene locus are shown for each strain and the resulting sequence type (ST). The *inlC* MVLST locus could not be found in all four lineage III strains (L99-4a, M7, HCC23, J2-071) and was found to be truncated by the end of a contig in F6854 and H7858.

Strain	*clpP*	*dal*	*inlB*	*inlC*	*lisR*	*prfA*	ST
02-1103	3	3	2	3	3	2	1
02-1289	3	3	2	3	3	2	1
02-1792	3	3	2	3	3	2	1
02-5993	1	1	7	1	1	1	2
02-6679	3	3	8	2	4	5	3
02-6680	3	3	8	2	4	5	3
04-5457	1	1	7	1	1	1	2
08-5578	1	1	7	1	1	1	2
08-5923	1	1	7	1	1	1	2
08-6056	1	1	7	1	1	1	2
08-6569	1	1	7	1	1	1	2
08-6997	1	1	7	1	1	1	2
08-7374	1	1	7	1	1	1	2
08-7669	1	1	7	1	1	1	2
10403S	2	4	15	1	1	3	4
10-0809	3	3	2	3	3	2	1
10-0810	3	5	3	2	4	4	5
10-0811	3	5	3	2	4	4	5
10-0812	2	4	4	1	1	3	6
10-0813	2	4	4	1	1	3	6
10-0814	1	1	7	1	1	1	2
10-0815	1	1	7	1	1	1	2
10-0933	2	2	4	1	1	1	7
10-0934	2	2	4	1	1	1	7
10-1046	1	1	7	1	1	1	2
10-1047	1	1	7	1	1	1	2
10-1321	1	1	7	1	1	1	2
10-4754	2	2	5	4	1	1	8

Continued on next page

Strain	*clpP*	*dal*	*inlB*	*inlC*	*lisR*	*prfA*	ST
10-4758	2	2	5	4	1	1	8
10-5024	1	1	7	1	1	1	2
10-5025	2	6	6	5	1	1	9
10-5026	2	6	6	5	1	1	9
10-5027	2	6	6	5	1	1	9
1988	2	1	7	1	1	1	10
81-0558	3	3	2	3	3	2	1
81-0592	3	3	2	3	3	2	1
81-0861	3	3	2	3	3	2	1
88-0478	1	1	7	1	1	1	2
95-0093	1	1	7	1	1	1	2
98-2035	1	1	7	1	1	1	2
99-6370	1	1	7	1	1	1	2
Clip80459	3	8	11	2	4	5	11
EGD-e	2	6	6	5	1	1	9
F2365	3	3	2	3	3	2	1
F6854	2	9	9	-	1	1	12
F6900	2	9	9	4	1	1	13
H7858	3	3	2	-	4	5	14
HCC23	4	11	10	-	6	6	15
HPB2262	3	5	8	3	3	5	16
J0161	2	9	9	4	1	1	13
J1-194	3	5	3	7	4	5	17
J2818	2	9	9	4	1	1	13
J2-071	5	12	14	-	8	8	18
L99-4a	4	11	10	-	6	6	15
M7	4	11	10	-	6	7	19
N1-017	3	5	12	2	3	5	20
N3-165	2	2	13	1	1	1	21
R2-503	3	5	12	2	3	5	20
R2-561	2	6	6	5	1	1	9
Scott-A	3	5	8	3	7	5	22