GENOME ANALYSIS OF *LISTERIA MONOCYTOGENES* AND *MELOSPIZA MELODIA*

by

SWARNALI LOUHA

(Under the Direction of Travis C. Glenn)

ABSTRACT

The advent of high throughput sequencing technologies makes it possible to address biological questions at the genome-wide scale. Analysis of these data creates unprecedented opportunities to explore the functions and dynamics of the genomes of large numbers of prokaryotes and non-model eukaryotic species. Herein I combine a new approach for allele-based subtyping and study of the genome of the pathogenic bacteria *Listeria monocytogenes*, and the analysis of the genome of the North American song sparrow (*Melospiza melodia*), selected for its behavioral, ecological, and biomedical importance.

I developed an open-source software (Haplo-ST) to provide whole-genome multi locus sequence typing (wgMLST) of *Listeria monocytogenes* from Illumina whole-genome sequencing data, while improving standardization and data exchangeability worldwide. Along with allelic profiles, this tool also generates allele sequences and identifies paralogous genes present in each isolate, which is extremely useful for evaluating phylogenetic relationships between closely related strains. More broadly, Haplo-ST is flexible and can be adapted to characterize the genome of any haploid organism simply by installing an organism-specific gene database. This tool was used to characterize and differentiate between two groups of *L. monocytogenes* isolates obtained from the natural environment and poultry processing plants. This tool was also used to study the patterns of genetic diversity and linkage disequilibrium in a large and diverse collection of *L. monocytogenes* isolates. We expect that Haplo-ST will serve as a valuable resource for accurately subtyping and evaluating relationships among bacterial isolates for routine surveillance, outbreak investigations and source tracking.

We used genome assembly and annotation of the North American song sparrow (*Melospiza melodia*) to identify genomic coordinates of protein-coding genes, microsatellites, repeat elements, transposable elements and several categories of non-coding RNA. The protein-coding genes were assigned with functional annotations and the genome assembly of the song sparrow was compared to that of several closely related birds. The genomic resources developed during this study will serve as valuable resources for facilitating studies contributing to biomedical research and in population genomic and comparative genomic studies of closely related species.

INDEX WORDS: *Listeria monocytogenes*, wgMLST, linkage disequilibrium, phylogeny, genetic variation, *Melospiza melodia*, reference genome, protein-coding genes, microsatellites, ncRNA, transposable elements, Haplo-ST, natural environment, poultry processing plants

GENOME ANALYSIS OF *LISTERIA MONOCYTOGENES* AND *MELOSPIZA MELODIA*

by

SWARNALI LOUHA

B. Sc., Bangalore University, India, 2007

M. Sc., Bangalore University, India, 2009

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

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2020

GENOME ANALYSIS OF *LISTERIA MONOCYTOGENES* AND *MELOSPIZA MELODIA*

by

SWARNALI LOUHA

Major Professor:	Travis C. Glenn
Committee:	Richard J. Meinersmann Zaid Abdo James H. Leebens-Mack

Electronic Version Approved:

Ron Walcott
Dean of the Graduate School
The University of Georgia
December 2020

DEDICATION

I dedicate this dissertation to my parents and my husband for their love, support, and inspiration to achieve my goals.

ACKNOWLEDGEMENTS

First of all, I am grateful to my advisor, Dr. Travis Glenn, for his consistent support and guidance to pursue the research described in this dissertation. From the very first day I rotated in his lab, he mentored and helped me develop a deep understanding of genomic technologies. Travis has always been there to help me and provided the best resources to explore areas away from his field of expertise. I am particularly indebted to him for exposing me to a variety of projects and providing me with the freedom to work on a wide range of ideas. I am also deeply indebted to Dr. Rick Meinersmann at USDA, Athens, for financially supporting me for the first few years at UGA and guiding me to learn the fundamentals of microbial population genetics. I would also like to thank Dr. Zaid and Dr. Leebens-Mack for providing valuable insights and comments and serving on my dissertation committee.

I also want to take this opportunity to thank Elizabeth, Hongye, Tito, Ruth, other IOB graduate students and members of the BadDNA lab with whom I interacted at both personal and professional fronts. We discussed many research ideas, participated in retreats, tried out different cuisines and did other fun stuff.

Finally, I would like to thank my family for their love and encouragement. My parents, who taught me the importance of education and gave me strength during difficult times, and my husband for the support and sacrifices he made to help me succeed in my endeavors. I am also thankful to my sister, who encouraged me and my two nephews for bringing joy in my life.

TABLE OF CONTENTS

Page

CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

1.1 GENOME CHARACTERIZATION OF *LISTERIA MONOCYTOGENES*

1.1.1 The foodborne pathogen, *Listeria monocytogenes*:

Listeria monocytogenes is a gram positive facultatively intracellular foodborne pathogen associated with significant morbidity and mortality worldwide, with an estimated 1600 cases of illnesses in the United States annually, resulting in more than 200 deaths. *Listeria monocytogenes* was first described by E. G. Murray and colleagues in 1926 (Murray et al. 1926). Although infection caused by *L. monocytogenes* was clinically described by the 1920s, it was not until 1952 that this organism was significantly associated with neonatal infection, sepsis and meningitis (Potel 1952). This pathogen was first identified to be a cause of foodborne illness in 1981 (Schlech et al. 1983), and later associated with infection in adults with compromised immune systems (Schlech 2000).

In humans, *L. monocytogenes* primarily causes a serious infection called listeriosis, which predominantly sickens people with weakened immune systems like pregnant woman, newborns, adults aged 65 and older, and those suffering from cancer, leukemia or transplant patients. Other than listeriosis, this pathogen also causes septicemia, encephalitis and meningitis in both children and adults with weak immune systems. Infection in healthy individuals has also been observed, although this is extremely rare (Vijila et al. 2007). In the United States, *L. monocytogenes* ranks third in fatality rates among foodborne bacterial pathogens, with 20-30% of deaths in high-risk individuals. In the European Union, listeriosis accounts for the highest proportion of hospitalized cases and deaths, making it one of the most serious foodborne diseases.

Listeria monocytogenes is well adapted as a saprophyte and ubiquitously present in the natural environment such as water, soil, and vegetation. Hence, it is easily contracted and transmitted by herd animals. Transmission may also arise from sources such as raw fruits and vegetables contaminated by environmental sources, direct or indirect contact with treated and untreated sewage, effluents from poultry and meat processing plants, decaying cereals like corn and soybeans and improperly fermented silage (Schuchat et al. 1992, Lorber & Bennett 2000, Henri et al. 2016). Further, *Listeria* is strongly adaptable to cold, acid, alkaline and osmotic stress, which enables it to grow in diverse environments (Raengpradub et al. 2008, Sue et al. 2004). Because of its ability to thrive at temperatures used for refrigeration (below 4°C), *L. monocytogenes* may also be transmitted by ready-to-eat foods such as unpasteurized milk, meat, poultry, seafood, fish and dairy products that are contaminated during manufacture, post-processing or storage in food facilities. Nearly all sporadic and epidemic human listeriosis cases have been linked to contaminated food or feed (Hyden et al. 2016). *Listeria* can also live in the intestines of birds, animals and humans for long periods of time without causing infection and has been found to be part of the normal gut microbiome of ~2-4% of healthy asymptomatic adults, most likely caused by the widespread agricultural carriage of this pathogen (Esteben et al. 2009, Gahan & Hill 2014). Within the host organism, quorum sensing and availability of nutrient resources help *Listeria* up-regulate the expression of virulence genes and acquire enhanced pathogenicity to cause severe infection (Garnet et al. 2006, Haber et al. 2017).

1.1.2 Genetic structure of *Listeria monocytogenes*

Listeria monocytogenes is generally considered to have a clonal genetic structure (Rasmussen et al. 1995, Wiedmann et al. 1997). Genome sequencing studies have shown that *L. monocytogenes* genomes are highly syntenic in nature (Kuenne et al. 2013) with a high degree of linkage disequilibrium existing between them (Call et al. 2003, Salcedo et al. 2003). Although the species pan-genome is highly stable, it is open to limited integration of foreign DNA, and evolutionary changes caused by mutation, duplication and recombination (Kuenne et al. 2013). Recombination observed between isolates belonging to different lineages of *L. monocytogenes* confirms that this species is not strictly clonal (den Bakker et al. 2008, Dunn et al. 2009, Ragon et al. 2008). Genetic diversity in this species is more likely to be driven by mutation than recombination (Ragon et al. 2008). Homologous recombination, which is rare, mostly occurs via conjugation and generalized transduction (Flamm et al. 1984, Lebrun et al. 1992, Hodgson 2000). Further, homologous recombination does not occur uniformly throughout the genome, but is more frequent in the accessory genome (Nelson et al. 2004, Hain et al. 2007, den Bakker et al. 2010).

On the basis of somatic (O) and flagellar (H) antigens, a total of 13 serotypes has been described in *L. monocytogenes*, with the majority of food-borne strains belonging to 1/2a, 1/2b, 1/2c and 4b, and serotypes 1a, 1b and 4b accounting for more than 90% of clinical isolates (Vijila et al. 2007, Henri et al. 2016). This species has also been divided into four distinct lineages: I, II, III and IV, each of which have distinct evolutionary histories, ecology, genomic content, recombination rates and pathogenic potential (Orsi et al. 2011, Haase et al. 2014). Each lineage is comprised of multiple serotypes; with lineage I containing serotypes 1/2b, 3b, 4b, 4d, 4e and 7; lineage II, serotypes 1/2a, 1/2c, 3a, 3c; lineage III: serotypes 1/2a, 4a, 4b and 4c; and lineage IV: 4a and 4c. About 96% of all human listeriosis cases are caused by Lineage I and II; serotypes 1/2a, 1/2b and 4b (7). Lineage I is more frequently associated with human listeriosis whereas lineage II strains are more commonly associated with food contamination and the environment and overrepresented in animal cases. Lineage III and IV strains occur less frequently among humans and have been linked to animal listeriosis (Dreyer et al. 2016).

1.1.3 Molecular subtyping of bacteria

The term 'subtype' is typically used to define groups below the level of bacterial species. A bacterial subtype is a group of organisms with the same attributes within a larger type. Subtyping methods identify common attributes which assign isolates to a larger type and different attributes that distinguish them from other subtypes (Bauer et al. 2013). Bacterial epidemiological typing aims to generate isolate-specific genotypic or phenotypic attributes that can be used to trace the sources and routes of bacterial dissemination. The scope of typing studies may vary from purely 'clinical' (transmission of infection from infected individuals or other sources to uninfected individuals) to 'environmental' (spread of microbes in inanimate surroundings) or 'industrial' (identification of microbes that are either valuable or a menace to the bio-industry) (van Belkum et al. 2007).

Molecular subtyping methods serve as valuable tools for the study of infectious disease pathogenesis, epidemiological surveillance and outbreak investigations conducted by public health agencies, and for tracking sources of microbial contaminants in the food processing industry. Typing may also be used for identifying emerging pathogenic strains, including potential agents of bioterrorism, or as evidence in forensic biology. In addition, molecular subtyping can also reveal markers of diversity contributing to bacterial population genetics, population structures and ecology.

1.1.4 Evolution of bacterial subtyping methods

In the history of classical and molecular microbiology, multiple procedures have been used for subtyping bacteria. Conventional typing methods such as bacteriophage typing, biochemical procedures, separation methods (SDS-PAGE, multi locus enzyme electrophoresis (MLEE), mass spectrometry) and serotyping have contributed towards understanding the natural history and epidemiology of infections caused by several clinically relevant bacterial pathogens (Wentworth 1963, Audurier and Martin 1989, Wolf 1997, Uzzau et al. 2000). In the field of clinical microbiology, antibiogram typing (antimicrobial susceptibility testing) has been used as a primary method for identification of bacterial cross-transmission in healthcare settings for a long time (van Belkum et al. 2007).

While bacterial phenotyping methods are helpful for elucidating healthcare associated outbreaks, they are limited in determining definitive relationships between isolates obtained from similar environments. Further, the development, application and quality control of some methods like phage typing and serotyping is costly, labor-intensive and require skills and methodologies that are difficult to maintain standards of today's accreditation bodies for microbiology laboratories. Additionally, because a given phenotype does not always accurately reflect the genotype of a microbe, phenotypic markers are unsuitable as stable epidemiological markers for critical endeavors like infection control and surveillance. Because of these limitations, phenotyping has largely been replaced by genotyping or 'molecular' typing over the last few decades. Over time, these methods have evolved from methods with poor standardization and reproducibility to highly standardized and reproducible methods with high discriminatory power.

Molecular subtyping methods can be differentiated into two broad categories; fragment-based and sequence-based methods. Here, I will present a brief overview of the molecular typing methods used for subtyping *L. monocytogenes* together with their advantages, limitations and unresolved issues of the methods currently used.

1.1.4.1 Fragment-based molecular subtyping methods

There are several types of fragment-based subtyping methods. Some methods depend on the separation of PCR-amplified DNA fragments based on molecular size such as amplified fragment length polymorphism (AFLP), multi-locus variable number tandem repeat analysis (MLVA), PCR-restriction fragment length polymorphism (PCR-RFLP) and random amplification of polymorphic DNA (RAPD) (Bauer et al. 2013, van Belkum et al. 2007). Other more commonly used methods rely on enzymatic digestion of bacterial DNA fragments such as ribotyping and pulsed field gel electrophoresis (PFGE) (Wiedmann 2002). PFGE has been considered as a gold standard for a long time and has been used by the Centers for Disease Control and Prevention (CDC) and state health departments in a national network (PulseNet) to exchange DNA subtypes for isolates of *L. monocytogenes.* Though this technique yields a high amount of pattern diversity that provides good discriminatory power, the relatedness of patterns is not a true measure of relatedness between isolates and is often used only as a guide. Further, PFGE is cumbersome, labor-intensive, subjective and becomes difficult as more profiles are entered into PFGE databases (Bauer et al. 2013).

1.1.4.2 Sequence-based molecular subtyping methods

Sequence-based subtyping methods are based on characterizing differences in short DNA sequences (several genes, SNPs), or entire genomes. A major advantage of these approaches is that sequence data is considerably less ambiguous and easier to interpret than banding pattern-based data obtained from fragment-based approaches. Further, sequence data does not require pure cultured isolates because sequences can be directly obtained by PCR amplification from clinical samples. More importantly, sequence data allows reconstruction of ancestral relationships among isolates, thus providing insights into bacterial evolution, ecology and epidemiology (Moorman et al. 2010).

Multi locus sequence typing (MLST), the genotypic descendent of MLEE, is a widely accepted approach that sequences DNA of several genes (usually 5-10) to identify bacterial subtypes and determines genetic relatedness between bacterial isolates (van Belkum et al. 2007). In this method, every isolate is defined by a sequence type (ST), which usually consists of a combination of seven allelic profiles, each distinct for that particular isolate. Groups of STs sharing a minimum of six identical alleles along with an ST acting as the ‘central genotype’ forms clonal complexes (CCs), which are geographically and temporally widespread (Henri et al. 2016). Although MLST generally characterizes differences in housekeeping genes, it may also be used for typing differences in virulence genes. Further, the development of online typing databases (Institut Pateur MLST database available at https://bigsdb.pasteur.fr/) that store allelic profiles of bacterial isolates facilitates standardized subtyping and allows large-scale surveillance studies. However, due to the use of a few slowly evolving housekeeping genes, conventional MLST lacks the resolution necessary to differentiate between closely related isolates that have diverged over short timeframes.

With the easy and cheap availability of next generation sequencing datasets, whole genome sequencing (WGS) of bacterial isolates have become readily available, leading to the development of high-throughput genome-wide SNP-based genotyping. This technique involves mapping sequencing reads from bacterial isolates to a reference genome and subtyping them based on the presence of SNPs at defined nucleotide positions known to be variable within a population. Several high quality SNP pipelines currently used are specifically designed to assess differences among closely related isolates (Jagadeesan et al. 2019). These include pipelines developed by the CDC (Lyve-SET, Katz et al. 2017), US FDA (CFSAN, Davis et al. 2015) and Applied Maths (BioNumerics). Because SNP-based approaches require the use of a closely related reference genome to prevent misalignment of reads against the reference genome and misidentification of SNPs, multiple CC-specific genomes are generally used as references to perform SNP calling within ST or CC groups. As the choice of multiple closely related reference genomes lack a global consensus, standardization of SNP-based approaches among different laboratories becomes difficult (Henri et al. 2017, Pearce et al. 2018).

The advent of next generation sequencing has also led to the extension of the conventional 7-gene MLST to multiple loci across the whole genome, thus providing high-throughput and high-resolution genotyping. These gene-by-gene approaches either use loci present in the core genome and shared by most isolates in a given population (termed as cgMLST, Moura et al. 2016), or all loci present in the pan-genome of a species (termed as wgMLST, Jagadeesan et al. 2019). Several open-source allele calling algorithms have been developed for cgMLST of *L. monocytogenes* (Pightling et al. 2015, Chen et al. 2016, Moura et al. 2016). Open-source algorithms that can perform wgMLST of bacterial isolates are also available and include Genome profiler (Zhang et al. 2015) and chewBBACA (Silva et al. 2018).

Genome profiler performs *ad hoc* wgMLST analysis of a set of bacterial genomes, whereas chewBBACA can be used to create and validate core and whole-genome MLST schemas using an algorithm based on BLAST score ratios. However, both these tools do not use any centralized nomenclature for assigning allele types, and hence are difficult to standardize across laboratories. Both core-genome and whole-genome MLST analysis have been implemented in commercial software as well, particularly BioNumerics (Applied Maths) and RidomSeqSphere+ (Ridom GmbH). Both platforms offer standardized allele-calling based on validated allelic profiles available in public databases, as well as the possibility to develop specific customized schemas. The US CDC has also created a wgMLST schema for *L. monocytogenes* inside PulseNet (using BioNumerics v7.5) making it feasible for federal, state and local public health laboratories to identify closely related isolates (Jackson et al. 2016). Although cgMLST/wgMLST based approaches are limited in that they are gene-centric and do not characterize differences in intergenic regions, their greatest advantage is that they are easy to standardize by using a unified allelic nomenclature.

1.1.5 Applications of molecular subtyping approaches to *L. monocytogenes*

Various molecular subtyping approaches have significantly improved our understanding of the biology, ecology and epidemiology of *L. monocytogenes* and other bacterial pathogens. In principle, the goal of molecular subtyping methods is to compare the genetic material of two or more bacterial isolates to determine whether they share a recent common ancestor. This has led to their application in surveillance programs which help in the rapid detection of foodborne disease outbreaks. Analysis of molecular subtyping data from human patients not only help in detecting widespread clusters of human foodborne disease cases, but also help in identifying and eliminating the outbreak source. Surveillance of human listeriosis cases is more challenging when compared to other foodborne pathogens (like *Salmonella* or *E. coli*) because *L. monocytogenes* has a long incubation period (7-60 days) and clinical disease develop in only specific sections of the population (newborn, elderly, pregnant, immunocompromised). Thus, effective surveillance and control of *L. monocytogenes* not only require sensitive subtyping approaches, but should also be accompanied with epidemiological data and a thorough understanding of bacterial genetics, population structure and physiology (Wiedman 2002).

Subtyping methods are also valuable in tracking the source of contamination in the food chain. *Listeria monocytogenes* has specifically been used as a model system for evaluating the efficiency of subtyping techniques in tracking in-plant *Listeria* contamination patterns (Wiedman 2002). This is because in contrast to other foodborne pathogens, *L. monocytogenes* can thrive in adverse conditions like cold, acid or alkaline environments within food processing facilities and form biofilms in food contact surfaces (Hyden et al. 2016). This gives rise to highly persistent strains that are difficult to remove with regular sanitization shifts and capable of re-contaminating the food processing environments multiple times. Thus, molecular subtyping methods help in controlling the spread of bacterial contamination and spoilage in the food industry.

Molecular subtyping approaches also provide an opportunity to explore the population genetics and evolution of *L. monocytogenes.* Subtyping methods help in defining *L. monocytogenes* isolates into subtypes and clonal groups, and associating them with phenotypic characteristics and pathogenic potential (Wiedman 2002). This can help recognize markers of pathogenicity in *L. monocytogenes* and estimate the virulence potential of strains isolated from infected individuals and contaminated food.

1.1.6 My objectives:

This dissertation applies development of an open-source bioinformatic approach for characterization of isolates of *L. monocytogenes,* and understanding patterns of linkage disequilibrium in *L. monocytogenes.*

In the initial work (chapter 2), computational approaches were used to develop a freely available and portable tool, Haplo-ST, that can perform wgMLST-based characterization of isolates of *L. monocytogenes* from short-read sequencing data, while allowing for data standardization and exchangeability worldwide. This tool was subsequently used for subtyping two groups of *L. monocytogenes* strains collected from different ecological niches (natural environment and poultry processing plants) and phylogenetic relationships were evaluated within members of each group. The phylogenetic analysis revealed clear delineation of isolates into lineages within each group and lineage-specificity was not observed with isolate origins or phenotypes. Further, genetic differentiation analysis was conducted within both groups of isolates and this revealed 21 highly differentiated loci in *L. monocytogenes* that were potentially enriched for adaptation and persistence of *L. monocytogenes* within poultry processing plants.

Haplo-ST was further used to characterize a diverse collection of 180 *L. monocytogenes* isolates from different geographical and temporal origins (chapter 3). This subtyping data was used for evaluating genetic variation and patterns of linkage disequilibrium in the pan-genome of *L. monocytogenes.* This analysis showed presence of strong linkage disequilibrium within the majority of genes in the genome of this bacteria. A set of 27 genes were found to have low levels of association with other genes and considered as potential hot spots for recombination events.

1.1.7 References

Audurier A, Martin C. 1989. Phage typing of *Listeria monocytogenes*. Int J Food Microbiol 8:251-257.

Bauer N, Evans P, Leopold B, Levine J, White P. 2013. USDA-FSIS Subtyping Work Group: Current and Future Development and Use of Molecular Subtyping by USDA-FSIS. Available at: https://www.fsis.usda.gov/wps/wcm/connect/6c7f71fd-2c0c-4ff0-b2bc-4977c7947516/Molecular-Subtyping-White-Paper.pdf?MOD=AJPERES

Call DR, Borucki MK, Besser TE. 2003. Mixed-genome Microarrays Reveal Multiple Serotype and Lineage-Specific Differences Among Strains of *Listeria monocytogenes*. J Clin Microbiol 41:632-639.

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

den Bakker HC, Cummings CA, Ferreira V, Vatta P, Orsi RH, Degoricija L, Barker M, Petrauskene O, Furtado MR, Wiedmann M. 2010. Comparative genomics of the bacterial genus *Listeria*: Genome evolution is characterized by limited gene acquisition and limited gene loss. BMC Genomics 11:688.

den Bakker HC, Didelot X, Fortes ED, Nightingale KK, Wiedmann M. 2008. Lineage specific recombination rates and microevolution in *Listeria monocytogenes*. BMC Evol Biol 8:277.

Dreyer M, Aguilar-Bultet L, Rupp S, Guldimann C, Stephan R, Schock A, Otter A, Schüpbach G, Brisse S, Lecuit M, Frey J, Oevermann A. 2016. *Listeria monocytogenes* sequence type 1 is predominant in ruminant rhombencephalitis. Sci Rep 6:36419.

Dunn KA, Bielawski JP, Ward TJ, Urquhart C, Gu H. 2009. Reconciling ecological and genomic divergence among lineages of Listeria under an “extended mosaic genome concept”. Mol Biol Evol 26:2605-2615.

Esteban JI, Oporto B, Aduriz G, Juste RA, Hurtado A. 2009. Faecal shedding and strain diversity of *Listeria monocytogenes* in healthy ruminants and swine in Northern Spain. BMC Vet Res 5:2.

Flamm RK, Hinrichs DJ, Thomashow MF. 1984. Introduction of pAM beta 1 into *Listeria monocytogenes* by conjugation and homology between native *L. monocytogenes* plasmids. Infect Immun 44:157-161.

Gahan CG, Hill C. 2014. *Listeria monocytogenes*: survival and adaptation in the gastrointestinal tract. Front Cell Infect Microbiol 4:9.

Garner MR, Njaa BL, Wiedmann M, Boor KJ. 2006. Sigma B contributes to *Listeria monocytogenes* gastrointestinal infection but not to systemic spread in the guinea pig infection model. Infect Immun 74:876-86.

Haase JK, Didelot X, Lecuit M, Korkeala H, *L. monocytogenes* MLST Study Group, Achtman M. 2014. The ubiquitous nature of *Listeria monocytogenes* clones: a large-scale Multilocus Sequence Typing study. Environ Microbiol 16:405-16.

Haber A, Friedman S, Lobel L, Burg-Golani T, Sigal N, Rose J, Livnat-Levanon N, Lewinson O, Herskovits AA. 2017. L-glutamine Induces Expression of *Listeria monocytogenes* Virulence Genes. PLoS Pathog 13:e1006161.

Hain T, Chatterjee SS, Ghai R, Kuenne CT, Billion A, Steinweg C, Domann E, Kärst U, Jänsch L, Wehland J, Eisenreich W, Bacher A, Joseph B, Schär J, Kreft J, Klumpp J, Loessner MJ, Dorscht J, Neuhaus K, Fuchs TM, Scherer S, Doumith M, Jacquet C, Martin P, Cossart P, Rusniock C, Glaser P, Buchrieser C, Goebel W, Chakraborty T. 2007. Pathogenomics of *Listeria* spp. Int J Med Microbiol 297:541-557.

Henri C, Félix B, Guillier L, Leekitcharoenphon P, Michelon D, Mariet JF, Aarestrup FM, Mistou MY, Hendriksen RS, Roussel S. 2016. Population Genetic Structure of *Listeria monocytogenes* Strains as Determined by Pulsed-Field Gel Electrophoresis and Multilocus Sequence Typing. Appl Environ Microbiol 82:5720-8.

Henri C, Leekitcharoenphon P, Carleton HA, Radomski N, Kaas RS, Mariet JF, Felten A, Aarestrup FM, Gerner Smidt P, Roussel S, Guillier L, Mistou MY, Hendriksen RS. 2017. An Assessment of Different Genomic Approaches for Inferring Phylogeny of *Listeria monocytogenes*. Front Microbiol 8:2351.

Hodgson DA. 2000. Generalized transduction of serotype 1/2 and serotype 4b strains of *Listeria monocytogenes*. Mol Microbiol 35:312-323.

Hyden P, Pietzka A, Lennkh A, Murer A, Springer B, Blaschitz M, Indra A, Huhulescu S, Allerberger F, Ruppitsch W, Sensen CW. 2016. Whole genome sequence-based serogrouping of *Listeria monocytogenes* isolates. J Biotechnol 235:181-6.

Jackson BR, Tarr C, Strain E, Jackson KA, Conrad A, Carleton H, Katz LS, Stroika S, Gould LH, Mody RK, Silk BJ, Beal J, Chen Y, Timme R, Doyle M, Fields A, Wise M, Tillman G, Defibaugh-Chavez S, Kucerova Z, Sabol A, Roache K, Trees E, Simmons M, Wasilenko J, Kubota K, Pouseele H, Klimke W, Besser J, Brown E, Allard M, Gerner-Smidt P. 2016. Implementation of nationwide real-time whole-genome sequencing to enhance listeriosis outbreak detection and investigation. Clin Infect Dis 63:380-6.

Jagadeesan B, Baert L, Wiedmann M, Orsi RH. 2019. Comparative Analysis of Tools and Approaches for Source Tracking *Listeria monocytogenes* in a Food Facility Using Whole-Genome Sequence Data. Front Microbiol 10:947.

Kuenne C, Billion A, Mraheil MA, Strittmatter A, Daniel R, Goesmann A, Barbuddhe S, Hain T, Chakraborty T. 2013. Reassessment of the *Listeria monocytogenes* Pan-Genome

Reveals Dynamic Integration Hotspots and Mobile Genetic Elements as Major Components of the Accessory Genome. BMC Genomics 14:47.

Lebrun M, Loulergue J, Chaslus-Dancla E, Audurier A. 1992. Plasmids in *Listeria monocytogenes* in relation to cadmium resistance. Appl Environ Microbiol 58:3183-3186.

Lorber B. 2000. *Listeria monocytogenes,* p 2208-14. *In* Mandell, Bennett, Dolan (ed), Mandell, Douglas, and Bennett's Principles and practice of infectious diseases, 5th ed.

Moorman M, Pruett P, Weidman M. 2010. Value and Methods for Molecular Subtyping of Bacteria, p 157-174. *In* Kornacki JL (ed), Principles of Microbiological Troubleshooting in the Industrial Food Processing Environment, 1st ed, Springer Science Business Media, New York, NY.

Moura A, Criscuolo A, Pouseele H, Maury MM, Leclercq A, Tarr C, Björkman JT, Dallman T, Reimer A, Enouf V, Larsonneur E, Carleton H, Bracq-Dieye H, Katz LS, Jones L, Touchon M, Tourdjman M, Walker M, Stroika S, Cantinelli T, Chenal-Francisque V, Kucerova Z, Rocha EPC, Nadon C, Grant K, Nielsen EM, Pot B, Gerner-Smidt P, Lecuit M, Brisse S. 2016. Whole genome-based population biology and epidemiological surveillance of *Listeria monocytogenes*. Nat Microbiol 2:16185.

Murray EG, Webb RA, Swann MB. 1926. A disease of rabbits characterized by a large mononuclear leucocytosis, caused by a hitherto undescribed bacillus Bacterium monocytogenes (n. sp.). J Pathol Bacteriol 29:407-439.

Nelson KE, Fouts DE, Mongodin EF, Ravel J, DeBoy RT, Kolonay JF, Rasko DA, Angiuoli SV, Gill SR, Paulsen IT, Peterson J, White O, Nelson WC, Nierman W, Beanan MJ, Brinkac LM, Daugherty SC, Dodson RJ, Durkin AS, Madupu R, Haft DH, Selengut J, Van Aken S, Khouri H, Fedorova N, Forberger H, Tran B, Kathariou S, Wonderling LD, Uhlich GA, Bayles DO, Luchansky JB, Fraser CM. 2004. Whole Genome Comparisons of Serotype 4b and 1/2a Strains of the Food-Borne Pathogen *Listeria monocytogenes* Reveal New Insights Into the Core Genome Components of This Species. Nucleic Acids Res 32:2386-2395.

Orsi RH, Bakker den HC, Wiedmann M. 2011. *Listeria monocytogenes* lineages: Genomics, evolution, ecology, and phenotypic characteristics. Int J Med Microbiol 301: 79-96.

Pearce ME, Alikhan N, Dallman TJ, Zhou Z, Grant K, Maiden MCJ. 2018. Comparative analysis of core genome MLST and SNP typing within a European Salmonella serovar Enteritidis outbreak. Int J Food Microbiol 274:1-11.

Pightling AW, Petronella N, Pagotto F. 2015. The *Listeria monocytogenes* core -genome sequence typer (LmCGST): a bioinformatics pipeline for molecular characterization with next generation sequence data. BMC Microbiol 15:224.

Potel J. 1952. Zur Granulomatosis infantiseptica. Zentr Bakteriol I Orig 158:329-331.

Raengpradub S, Wiedmann M, Boor KJ. 2008. Comparative analysis of the sigma B-dependent stress responses in *Listeria monocytogenes* and *Listeria innocua* strains exposed to selected stress conditions. Appl Environ Microbiol 74:158-171.

Ragon M, Wirth T, Hollandt F, Lavenir R, Lecuit M, Le Monnier A, Brisse S. 2008. A new perspective on *Listeria monocytogenes* evolution. PLoS Pathog 4:e1000146.

Ramaswamy V, Cresence VM, Rejitha JS, Lekshmi MU, Dharsana KS, Prasad SP, Vijila HM. 2007. *Listeria*--review of epidemiology and pathogenesis. J Microbiol Immunol Infect 40:4-13.

Rasmussen OF, Skouboe P, Dons L, Rossen L, Olsen JE. 1995. *Listeria monocytogenes* exists in at least three evolutionary lines: evidence from flagellin, invasive associated protein and listeriolysin O genes. Microbiology 141:2053-2061.

Salcedo C, Arreaza L, Alcalá B, de la Fuente L, Vázquez JA. 2003. Development of a multilocus sequence typing method for the analysis of Listeria monocytogenes clones. J Clin Microbiol 41:757-762.

Schlech WF, Lavigne PM, Bortolussi RA, Allen AC, Haldane EV, Wort AJ, Hightower AW, Johnson SE, King SH, Nicholls ES, Broome CV. 1983. Epidemic listeriosis--evidence for transmission by food. N Engl J Med 308:203-6.

Schlech WF. 2000. Foodborne listeriosis. Clin Infect Dis 31:770-5.

Schuchat A, Deaver KA, Wenger JD, Plikaytis BD, Mascola L, Pinner RW, Reingold AL, Broome CV. 1992. Role of Foods in Sporadic Listeriosis. I. Case-control Study of Dietary Risk Factors. The *Listeria* Study Group. JAMA 267:2041-5.

Silva M, Machado M, Silva D, Rossi M, Moran-Gilad J, Santos S, Ramirez M, Carriço J. 2018. chewBBACA: A complete suite for gene-by-gene schema creation and strain identification. Microb Genom 4:e000166.

Sue D, Fink D, Wiedmann M, Boor KJ. 2004. Sigma B-dependent gene induction and expression in *Listeria monocytogenes* during osmotic and acid stress conditions simulating the intestinal environment. Microbiology 150:3843-4455.

Uzzau S, Brown DJ, Wallis T, Rubino S, Leori G, Bernard S, Casadesús J, Platt DJ, Olsen JE. 2000. Host adapted serotypes of *Salmonella enterica*. Epidemiol Infect 125:229-255.

van Belkum A, Tassios PT, Dijkshoorn L, Haeggman S, Cookson B, Fry NK, Fussing V, Green J, Feil E, Gerner-Smidt P, Brisse S, Struelens M; European Society of Clinical Microbiology and Infectious Diseases (ESCMID) Study Group on Epidemiological Markers (ESGEM). 2007. Guidelines for the validation and application of typing methods for use in bacterial epidemiology. Clin Microbiol 3:1-46.

Wentworth BB. 1963. Bacteriophage typing of the Staphylococci. Bacteriol Rev 27:253-272.

Wiedmann M, Bruce JL, Keating C, Johnson AE, McDonough PL, Batt CA. 1997. Ribotypes and virulence gene polymorphisms suggest three distinct *Listeria monocytogenes* lineages with differences in pathogenic potential. Infect Immun 65:2707-2716.

Wiedmann M. 2002. Molecular subtyping methods for *Listeria monocytogenes*. J AOAC Int 85:524-531.

Wolf MK. 1997. Occurrence, distribution and associations of O and H serogroups, colonization factor antigens and toxins of enterotoxigenic *Escherichia coli*. Clin Microbiol Rev 10:569-584.

Zhang J, Halkilahti J, Hänninen M, Rossi M. 2015. Refinement of whole-genome multilocus sequence typing analysis by addressing gene paralogy. J Clin Microbiol 53:1765-7.

1.2 SECTION II: GENOME ANALYSIS OF *MELOSPIZA MELODIA*

1.2.1 Genome analysis

A genome is an organism's complete set of DNA, including all of its genes and inter-genic regions, which contains all the information needed to build and maintain the organism. Identifying and quantifying all of an organism's genes and their interactions with each other as well as the environment can unravel their functions and consequent effects on the organism. Therefore, analysis of genomes is essential for understanding the genetic information written in the DNA of an organism. Genome analysis also includes DNA sequencing, assembly of DNA to represent original chromosomes and analysis of the resulting assembly for structure and function (Pevsner 2009). Thus, a genome provides valuable shortcuts, helping researchers find genes and other non-genic feature of interest easily and quickly. Further, the study of genomes also involves the study of intragenomic processes such as epistasis, heterosis and pleiotropy. Genome analysis has been a key area of biological investigation for decades. Research in this field has progressed from Sanger sequencing using radiolabeled primers to early shotgun sequencing with bacterial vectors to high-throughput sequencing using second and third generation sequencing technologies (Giani et al. 2020). Here, I will present a brief history of the rise and evolution of genome research and its applications, and then provide the objectives for analyzing the genome of the North American song sparrow (*Melospiza melodia*), which has been considered as a model vertebrate species in field studies of birds.

1.2.2 First generation sequencing

DNA was first identified in 1869 by Friedrich Miescher. However, it took over a century to improve understanding of the nature of DNA, including the nucleotide bases that compose it and the theory of chromosomal inheritance. It was during this period that Johannsen introduced the concept of ‘gene’ and Hans Winkler first proposed the term ‘genome’ to designate the complete genetic makeup of an organism (Weissenbach 2016). The double helical structure of DNA and the “codon for life” that guides the production of specific proteins were also discovered during this period. Technology continued to progress and RNA sequencing became feasible in the late 1960s with transfer and ribosomal RNAs to be the first RNA molecules to be decoded (Holley et al 1965, Brownlee et al. 1968). The real progression in DNA sequencing was set into motion in 1975 when the ‘plus and minus’ method was developed and used to sequence two short regions in the genome of phage φX174 (Sanger and Coulson 1975). Two years later, the same approach was used by Fred Sanger to sequence the first DNA genome of phage φX174 (Sanger et al. 1977a). In the same year, Sanger developed a new method that could decipher DNA fragments of approximately 400 bases in a day. This classical method, known as ‘Sanger sequencing’ is based on selective incorporation of chain-terminating dideoxynucleotides by DNA polymerase during *in vitro* DNA replication. This involves four sets of polymerization reactions using tritium-radiolabeled primers, where each reaction is supplied with small amounts of one chain-terminating 2,3-dideoxynucleoside triphosphate (ddNTP) to produce fragments of different lengths (Sanger et al. 1977b). When DNA polymerase incorporates a ddNTP at the 3' end of the growing DNA strand, chain elongation is terminated due to a missing 3' hydroxyl group (Atkinson et al. 1969). The products of the four reactions are loaded on polyacrylamide gels and the sequence is deduced by comparing the size of the fragments. This method was widely used for approximately 30 years, after which it was replaced by high-throughput sequencing platforms.

1.2.3 Shotgun sequencing

In an attempt to accelerate DNA sequencing, Staden proposed "Shotgun sequencing" in 1979, in which bacterial vectors are used to clone random fragments of a long DNA molecule, which are sequenced in parallel and assembled using read overlaps (Staden 1979). This approach was used by Messing to develop the first shotgun sequencing protocol (Messing et al. 1981) and later adopted by Sanger to assemble the 48,502 bp long genome of phage λ (Sanger et al. 1982). In the next two decades, many genome sequencing projects were launched and completed, leading to a large increase of data available in public repositories such as GenBank. Sequencing of entire genomes of a multitude of unicellular microorganisms (*Epstein-Barr* virus, *Vaccinia* virus, *Human cytomegalovirus*, *H. influenza*, *E. coli*) were undertaken followed by sequencing projects of more complex eukaryotes (*C. elegans*, *Arabidopsis*). During this time, advances in technological developments and industrial processes increased throughput and decreased sequencing errors. Major milestones included the synthesis of fluorescent DNA primers and their use in automating Sanger sequencing (Smith et al. 1986), introduction of dye terminator sequencing (Prober et al. 1987), and the release of the first commercial florescence automated DNA sequencer (ABI 370A) by Applied Biosystems. The introduction of highly optimized DNA polymerases further increased the speed and efficiency of sequencing (Tabor et al. 1987, Murray et al. 1989). These technological breakthroughs were followed by the introduction of 'bodipy' dyes that were more effective than conventional dyes (Metzker et al. 1996), magnetic bead-based DNA purification methods (DeAngelis et al. 1995), and capillary electrophoresis (Zhang et al. 1995). These initiatives were further accelerated by the Human Genome Project, which aimed to assemble the complete set of human chromosomes (Lander et al. 2001), and its competition with Celera, a private company that also tried to achieve the same goals independently (Venter et al.

2001). The race to assemble the human genome made commercial enterprises realize the potential of a profitable business in sequencing and prompted the development of a diverse range of sequencing technologies, collectively known as Next-Generation Sequencing (NGS).

1.2.4 Next-Generation Sequencing

In the 2000s, the enthusiasm generated by the Human Genome Project gave birth to many private companies offering a variety of sequencing technologies at higher throughput and lower costs. These included 454, Solexa, Agencourt, Illumina, Complete Genomics and Applied Biosystems (Giani et al. 2020). The first NGS sequencer, GS20, was based on pyrosequencing and commercialized by 454 Life Sciences. This technology starts with single molecule template synthesis of small bead-bound DNA fragments, which are amplified in a water-in-oil emulsion clonal PCR (Tawfik and Griffiths 1998). The beads are then loaded into picotitre plates and sequenced in parallel by flowing pyrosequencing reagents across the plate (Glenn 2011). This system could produce 400-500 bp reads, had a 99% accuracy and could sequence a maximum of 25 million bp in a 4 hr period at one-sixth the costs of conventional methods (Giani et al. 2020). 454 was later acquired by Roche but is still known by the name 454.

In the next few years, Solexa added several newer technologies (Kawashima et al. 1998, Mitra et al. 2003, Ruparel et al. 2005, Seo et al. 2005, Ost 2006) which further increased sequencing throughput and produced stronger optical signals. Solexa was acquired by Illumina and has established itself as one of the most popular sequencing platforms today. Illumina uses a glass flow cell for capturing and amplifying DNA fragments into clusters of identical molecules with a technique known as bridge amplification (Kawashima et al. 1998). The amplified clusters are sequenced with an approach similar to Sanger sequencing, except that dye labelled terminators are used to detect every single nucleotide added to the end of the growing DNA chain. The greatest advantage of Illumina at that time was it represented paired-end reads i.e., both strands in a DNA molecule could be sequenced (one, then the other), which allowed to gauze the gap size between distant sites on a DNA fragment.

In 2007, SOLiD, the 3rd commercial NGS technology, was introduced by Applied Biosystems. SOLiD uses the specificity of DNA ligases to determine sequences (Brenner et al. 2000), but this method was found to have issues with sequencing palindromic sequences and consequently abandoned. Around the same time, Helicos developed the first commercial single molecule sequencer (Braslavsky et al. 2003, Harris et al. 2008). However, the short read-lengths together with high costs of sequencing limited usage of this platform. Ion Torrent developed a sequencing technology in 2011 which measured the pH variations induced by the release of protons during DNA synthesis. The hydrogen ions released during nucleotide additions were detected using a semiconductor sensor (Rothberg et al. 2011). Although this technology is still in use, it suffers from inaccuracies in the measurement of homopolymers (Loman et al. 2012). The limitations of these platforms led the way to the commercial success of Illumina which appeared to have a near monopoly over the DNA sequencing market by 2014. The release of the HiSeq and NovaSeq instruments by Illumina has drastically increased the throughput and reduced sequencing costs in the last decade.

The Second-Generation Sequencing technologies mentioned above has allowed cost-effective and rapid resequencing of genomes together with novel applications such as RNAseq, ChIP-seq, whole exome sequencing, genotyping with SNPs and epigenetic landscape determination (Thorisson et al. 2005, Johnson et al. 2007, Lister et al. 2008, Nagalakshmi et al. 2008, Ng et al. 2009). However, difficulty in detecting overlaps between short-reads produced by these platforms led to partially complete draft genomes, with problems in attaining complete sequences of repeats and low complexity regions (Bailey et al. 2002, Alkan et al. 2011). Newer technologies offered by PacBio and Oxford Nanopore have attempted to fill this gap by opening up the era of Third-Generation Sequencing in the last decade. In contrast to the Second-Generation Sequencing, single DNA molecules are sequenced in nearly real-time and produce much longer reads spanning one to several hundred kilobases. The availability of ultralong reads greatly improves the quality of genome assemblies, as it enables generation of long continuous consensus sequences (Rhoades et al. 2015, Giordano et al. 2017). Although, in the past, the individual base-calling accuracy of these platforms were far less than Illumina reads, it has gradually increased over the years culminating in the release of a new method by PacBio in 2019 called 'HiFi' (High Fidelity), which can generate 10-20 kbp long reads that are as accurate as Illumina short reads (Wenger et al. 2019). The subsequent release of Sequel II by PacBio has further increased the throughput of sequencing to 160 Gb per SMRT Cell, with a concomitant drop of up to 8-fold in sequencing costs (Giani et al. 2020). On the other hand, reads produced by Nanopore (Huang et al. 2010, Cherf et al. 2012, Manrao et al. 2012) have been recorded to be longer than PacBio (Payne et al. 2019) and found to sequence through repeats where even PacBio reads may fail (Giani et al. 2020). Another advantage with Nanopore is that this platform provides devices as small as a USB stick, allowing easy portability in remote field sites (Quick et al. 2016). However, the Nanopore technology is also known to produce sequencing biases that are difficult to correct (Istace et al. 2017).

Other than providing improved genome assemblies with high structural accuracy, the Third-Generation Sequencing technologies allow generation of long phased blocks of haplotypes, where the paternal and maternal contributions to a homologous region of the chromosome are reported separately (Kuleshov et al. 2014). This facilitates accurate mapping of reads for structural variant detection, including length variations in repeat motifs, indels, duplications, inversions and translocations (Merker et al. 2018, Chaisson et al. 2019). Further, a variety of epigenetic markers can also be characterized along with DNA sequencing in both PacBio and Nanopore Sequencing technologies (Schadt et al. 2013, Rand et al. 2017).

1.2.5 Supporting technologies

Several supporting technologies are also used to improve the contiguity of existing genome assemblies. These include optical mapping platforms (e.g., Bio-Nano), linked-read technologies (e.g., 10X Genomics Chromium system), or the genome-folding approach of Hi-C from Dovetail Genomics. These technologies help in orienting contigs in their putative order on chromosomes, by a process known as scaffolding.

The current optical mapping technique used by BioNano focusses on labelling DNA molecules with specific restriction enzymes and imaging them with a high-resolution camera. Information from individual DNA molecules are combined to form consensus optical maps, which provide the linkage information needed to improve the process of *de novo* genome assembly (Teague et al. 2010, Lam et al. 2012), as well as identify and rectify misassemblies (Tang et al. 2015, Howe et al. 2015). Improved contiguity of hybrid genome assemblies also allows the detection of structural variants by comparing to a reference (Mak et al. 2016) and helps in identifying genome-wide methylation patterns through methylation-sensitive restriction enzymes (Ananiev et al. 2008). The linked read technology offered by 10X Genomics leverages microfluidics to partition and barcode high molecular weight DNA to generate linked reads, which provide long-range information from short-read sequencing data. However, this technology has been discontinued in 2020.

Another scaffolding technique called Hi-C uses chromatin proximity information for all regions of the genome to arrange contigs and scaffolds in a linear sequences of chromosomes (Lieberman-Aiden et al. 2009, Burton et al. 2013, Kaplan et al. 2013). In this technique, cells are embedded in a matrix and treated to remove all layers except the chromosome folding information. The DNA is then cleaved with restriction enzymes and re-ligated to form covalent bonds with new, spatially close molecules. The resulting library is sequenced with Illumina short reads and the DNA interactions obtained from the ligation step is used to order contigs into chromosomes (Dudchenko et al. 2017, Teh et al. 2017). While Hi-C has the potential to scaffold large genomes, it also produces misassembly errors such as artificial inversions, scaffold misplacement within the same chromosome, or scaffold misassignment to different chromosomes (Burton et al. 2013). To correct these errors, a combination of different scaffolding techniques have been adopted by many sequencing projects (Bickhart et al. 2017, Wallberg et al. 2019).

Ultimately, the choice of a specific sequencing technology depends on several factors such as the research goals, availability of a particular technology, amount of DNA available for sequencing, and associated costs. Often, a combination of both long and short read technologies is adopted, as shorter reads have a different error profile and can be used to correct longer ones (Koren et al. 2012).

1.2.6 Applications of genome analysis

Information obtained from genome analysis can be applied to a variety of fields including medicine, biotechnology, agriculture and social sciences. Here, I will briefly list a few major applications of genome research.

1.2.6.1 Annotation of biologically significant elements:

A raw genome sequence is not of much value unless it has been annotated for biologically meaningful information. This involves analyzing the sequence structure and composition, as well as using information from closely related reference species to determine a variety of biologically significant elements in the genome. While genomes are annotated for repetitive elements, microsatellites, non-coding RNAs etc., genome annotation projects mainly focus on correctly identifying the location, structure and function of protein-coding genes. Gene prediction has been aided with the development of many algorithms in the past decade. These approaches can be classified as 'intrinsic', 'extrinsic' and 'combiner' approaches (Del Angel et al. 2018). Intrinsic approaches consist of *ab-initio* gene prediction, where a training set of statistical models is used to extract information from the genomic sequence itself such as coding potential, splice site prediction etc. On the other hand, extrinsic approaches use sequences available in public repositories like transcripts, ESTs, RNA-seq for gene prediction. Both of these approaches have their own advantages and disadvantages, and 'combiner' approaches integrate the best of both techniques. Combiners like 'Eugene' and 'Maker' predict genes using an integrated approach in which intrinsic prediction is modified by a given extrinsic dataset. Thus, the quality of final results not only depend on the choice of algorithm, but also on the selection of the dataset provided to the algorithm. Finally, the most important step in gene prediction is to provide functional annotations to the predicted polypeptides by comparing their similarity to other sequences present in public repositories. Downstream analysis of the functional annotation process allows further understanding of specific genome properties such as metabolic pathways, similarity to closely related species etc. Genome sequencing and annotation also favors comparative genomics of closely related organisms and detection of selection.

1.2.6.2 Identification of variants:

A prominent application of genome sequencing is to identify variants from sequenced genomes for studying genetic association with diseases, detecting mutations in cancer, or characterizing heterogenous cell populations (Liu et al. 2013). These variants include single nucleotide polymorphisms (SNPs), single nucleotide variants (SNVs), copy number variations (CNVs) and structural variants (SVs), which can be used as genetic markers for screening diseases. Detection of SNVs and indels are essential for understanding the genetics of diseases and help in clinical diagnosis and treatment of patients (Altmann et al. 2012). CNVs play a role in human diversity and their impact on disease susceptibility has been well recorded (Pirooznia et al. 2015). With the increase in whole genome sequencing, there has been a dramatic increase in the number of variants and their complexity. This has prompted prediction of the functional impacts of variants and their implications in diseases. Variant calling is also used to study the amount of variation within and between populations in population genetic studies (Wright et al. 2019). In plant genetics, variants such as microsatellite repeats, simple sequence repeats and SNPs are used as genetic markers (Chaitanya 2019).

1.2.6.3 Detection of epigenetic markers:

Genome sequencing and analysis allows for the study of the epigenome of an organism along with post-translational modification of histones and methylation maps, both of which are regarded as epigenetic markers. DNA methylation is essential for normal development of an organism and plays a crucial role in many vital processes (Chaitanya 2019). However, hypomethylation and hypermethylation of CpG islands in specific regions of the genome leads to cancer. DNA methylation can be studied by methylated DNA immunoprecipitation (meDIP) (Thu et al. 2009). Modification of histone proteins can also affect DNA indirectly by altering regulation of gene expression. Post translational modifications of histones at the genome level can be identified with ChIP-Seq technology (Veluchamy et al. 2015).

1.2.7 My objectives:

I analyzed the genome of the North American song sparrow, *Melospiza melodia* (chapter 4), which has been widely studied for its behavioral and ecological characteristics, and is a favorable candidate in several areas of biomedical research. The primary objective of sequencing and analyzing the genome of *M. melodia* was to provide a reference genome assembly and its associated annotations for this species. To achieve these goals, a Chicago library of the genome of *M. melodia* was sequenced and assembled with the HiRise scaffolding software pipeline at Dovetail Genomics. The resulting genome assembly was annotated for protein coding genes and other non-genic features of interest such as transposable elements, microsatellites and non-coding RNAs. This study yielded in a high-quality and highly complete *de-novo* genome assembly of *M. melodia* which will serve as a reference for a variety of genetic, ecological, functional and comparative genomic studies in songbirds and other related taxa.

1.2.8 References:

Alkan C, Sajjadian S, Eichler EE. 2011. Limitations of next-generation genome sequence assembly. Nat Methods 8:61-5.

Altmann A, Weber P, Bader D, Preuss M, Binder EB, Müller-Myhsok B. 2012. A beginners guide to SNP calling from high-throughput DNA-sequencing data. Hum Genet 131:1541-1554.

Ananiev GE, Goldstein S, Runnheim R, Forrest DK, Zhou S, Potamousis K, Churas CP, Bergendahl V, Thomson JA, Schwartz DC. 2008. Optical mapping discerns genome wide DNA methylation profiles. BMC Mol Biol 9:68.

Atkinson MR, Deutscher MP, Kornberg A, Russell AF, Moffatt JG. 1969. Enzymatic synthesis of deoxyribonucleic acid. XXXIV. Termination of chain growth by a 2',3'-dideoxyribonucleotide. Biochemistry 8:4897-904.

Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, Schwartz S, Adams MD, Myers EW, Li PW, Eichler EE. 2002. Recent segmental duplications in the human genome. Science 297:1003-7.

Bickhart DM, Rosen BD, Koren S, Sayre BL, Hastie AR, Chan S, Lee J, Lam ET, Liachko I, Sullivan ST, Burton JN, Huson HJ, Nystrom JC, Kelley CM, Hutchison JL, Zhou Y, Sun J, Crisà A, Ponce de León FA, Schwartz JC, Hammond JA, Waldbieser GC, Schroeder SG, Liu GE, Dunham MJ, Shendure J, Sonstegard TS, Phillippy AM, Van Tassell CP, Smith TP. 2017. Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome. Nat Genet 49:643-50.

Braslavsky I, Hebert B, Kartalov E, Quake SR. 2003. Sequence information can be obtained from single DNA molecules. Proc Natl Acad Sci USA 100:3960-4.

Brenner S, Johnson M, Bridgham J, Golda G, Lloyd DH, Johnson D, Luo S, McCurdy S, Foy M, Ewan M, Roth R, George D, Eletr S, Albrecht G, Vermaas E, Williams SR, Moon K, Burcham T, Pallas M, DuBridge RB, Kirchner J, Fearon K, Mao J, Corcoran K. 2000. Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. Nat Biotechnol 18:630-4.

Brownlee GG, Sanger F, Barrell BG. 1968. The sequence of 5s ribosomal ribonucleic acid. J Mol Biol 34:379-412.

Burton JN, Adey A, Patwardhan RP, Qiu R, Kitzman JO, Shendure J. 2013. Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. Nat Biotechnol 31:1119-25.

Chaisson MJP, Sanders AD, Zhao X, Malhotra A, Porubsky D, et al. 2019. Multi-platform discovery of haplotype-resolved structural variation in human genomes. Nat Commun 10:1784.

Chaitanya KV. 2019. Applications of Genomics, p 243-260. *In* Genome and Genomics, Springer, Singapore.

Cherf GM, Lieberman KR, Rashid H, Lam CE, Karplus K, Akeson M. 2012. Automated forward and reverse ratcheting of DNA in a nanopore at 5-Å precision. Nat Biotechnol 30:344-8.

DeAngelis MM, Wang DG, Hawkins TL. 1995. Solid-phase reversible immobilization for the isolation of PCR products. Nucleic Acids Res 23:4742-3.

Dominguez Del Angel V, Hjerde E, Sterck L, Capella-Gutierrez S, Notredame C, Vinnere Pettersson O, Amselem J, Bouri L, Bocs S, Klopp C, Gibrat JF, Vlasova A, Leskosek BL, Soler L, Binzer-Panchal M, Lantz H. 2018. Ten steps to get started in Genome Assembly and Annotation. F1000Res 7:ELIXIR-148.

Dudchenko O, Batra SS, Omer AD, Nyquist SK, Hoeger M, Durand NC, Shamim MS, Machol I, Lander ES, Aiden AP, Aiden EL. 2017. De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. Science 356:92-5.

Giordano F, Aigrain L, Quail MA, Coupland P, Bonfield JK, Davies RM, Tischler G, Jackson DK, Keane TM, Li J, Yue JX, Liti G, Durbin R, Ning Z. 2017. De novo yeast genome assemblies from MinION, PacBio and MiSeq platforms. Sci Rep 7:3935.

Glenn TC. 2011. Field guide to next-generation DNA sequencers. Mol Ecol Resour 11:759-69.

Harris TD, Buzby PR, Babcock H, Beer E, Bowers J, Braslavsky I, Causey M, Colonell J, Dimeo J, Efcavitch JW, Giladi E, Gill J, Healy J, Jarosz M, Lapen D, Moulton K, Quake SR, Steinmann K, Thayer E, Tyurina A, Ward R, Weiss H, Xie Z. 2008. Single-molecule DNA sequencing of a viral genome. Science 320:106-9.

Holley RW, Apgar J, Everett GA, Madison JT, Marquisee M, Merrill SH, Penswick JR, Zamir A. 1965. Structure of a ribonucleic acid. Science 147:1462-5.

Howe K, Wood JMD. 2015. Using optical mapping data for the improvement of vertebrate genome assemblies. GigaScience 4:10.

Huang S, He J, Chang S, Zhang P, Liang F, Li S, Tuchband M, Fuhrmann A, Ros R, Lindsay S. 2010. Identifying single bases in a DNA oligomer with electron tunnelling. Nat Nanotechnol 5:868-73.

Istace B, Friedrich A, d'Agata L, Faye S, Payen E, Beluche O, Caradec C, Davidas S, Cruaud C, Liti G, Lemainque A, Engelen S, Wincker P, Schacherer J, Aury JM. 2017. de novo assembly and population genomic survey of natural yeast isolates with the Oxford Nanopore MinION sequencer. GigaScience 6:1-13.

Johnson DS, Mortazavi A, Myers RM, Wold B. 2007. Genome-wide mapping of in vivo protein-DNA interactions. Science 316:1497-502.

Kaplan N, Dekker J. 2013. High-throughput genome scaffolding from in vivo DNA interaction frequency. Nat Biotechnol 31:1143-7.

Kawashima E, Farinelli L, Mayer P. 1998. Method of nucleic acid amplification. WIPO Patent WO1998044151A1.

Koren S, Schatz MC, Walenz BP, Martin J, Howard JT, Ganapathy G, Wang Z, Rasko DA, McCombie WR, Jarvis ED, Adam M Phillippy. 2012. Hybrid error correction and de novo assembly of single-molecule sequencing reads. Nat Biotechnol 30:693-700.

Kuleshov V, Xie D, Chen R, Pushkarev D, Ma Z, Blauwkamp T, Kertesz M, Snyder M. 2014. Whole-genome haplotyping using long reads and statistical methods. Nat Biotechnol 32:261-6.

Lam ET, Hastie A, Lin C, Ehrlich D, Das SK, Austin MD, Deshpande P, Cao H, Nagarajan N, Xiao M, Kwok PY. 2012. Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. Nat Biotechnol 30:771-6.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, et al. 2001. Initial sequencing and analysis of the human genome. Nature 409:860-921.

Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, Sandstrom R, Bernstein B, Bender MA, Groudine M, Gnirke A, Stamatoyannopoulos J, Mirny LA, Lander ES, Dekker J. 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science 326:289-93.

Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, et al. 2008. Highly integrated single-base resolution maps of the epigenome in Arabidopsis. Cell 133:523-36.

Liu X, Han S, Wang Z, Gelernter J, Yang BZ. 2013. Variant callers for next-generation sequencing data: a comparison study. PLoS One 8:e75619.

Loman NJ, Misra RV, Dallman TJ, Constantinidou C, Gharbia SE, Wain J, Pallen MJ. 2012. Performance comparison of benchtop high-throughput sequencing platforms. Nat Biotechnol 30:434-9.

Mak AC, Lai YY, Lam ET, Kwok TP, Leung AK, Poon A, Mostovoy Y, Hastie AR, Stedman W, Anantharaman T, Andrews W, Zhou X, Pang AW, Dai H, Chu C, Lin C, Wu JJ, Li CM, Li JW, Yim AK, Chan S, Sibert J, Džakula Ž, Cao H, Yiu SM, Chan TF, Yip KY, Xiao M, Kwok PY. 2016. Genome-wide structural variation detection by genome mapping on nanochannel arrays. Genetics 202:351-62.

Manrao EA, Derrington IM, Laszlo AH, Langford KW, Hopper MK, Gillgren N, Pavlenok M, Niederweis M, Gundlach JH. 2012. Reading DNA at single-nucleotide resolution with a mutant MspA nanopore and phi29 DNA polymerase. Nat Biotechnol 30:349-53.

Merker JD, Wenger AM, Sneddon T, Grove M, Zappala Z, Fresard L, Waggott D, Utiramerur S, Hou Y, Smith KS, Montgomery SB, Wheeler M, Buchan JG, Lambert CC, Eng KS, Hickey L, Korlach J, Ford J, Ashley EA. 2018. Long-read genome sequencing identifies causal structural variation in a Mendelian disease. Genet Med 20:159-63.

Messing J, Crea R, Seeburg PH. 1981. A system for shotgun DNA sequencing. Nucleic Acids Res 9:309-21.

Metzker ML, Lu J, Gibbs RA. 1996. Electrophoretically uniform fluorescent dyes for automated DNA sequencing. Science 271:1420-2.

Mitra RD, Shendure J, Olejnik J, Edyta-Krzymanska-Olejnik, Church GM. 2003. Fluorescent in situ sequencing on polymerase colonies. Anal Biochem 320:55-65.

Murray V. 1989. Improved double-stranded DNA sequencing using the linear polymerase chain reaction. Nucleic Acids Res 17:8889.

Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, et al. 2008. The transcriptional landscape of the yeast genome defined by RNA sequencing. Science 320:1344-9.

Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, et al. 2009. Targeted capture and massively parallel sequencing of 12 human exomes. Nature 461:272-6.

Ost TB, Smith GP, Balasubramanian S, Rigatti R, Sanches RM. 2006. Improved polymerases. WIPO Patent WO2006120433.

Payne A, Holmes N, Rakyan V, Loose M. 2019. BulkVis: a graphical viewer for Oxford nanopore bulk FAST5 files. Bioinformatics 35:2193-8.

Pevsner J. 2009. Wiley-Blackwell (ed), Bioinformatics and functional genomics, 2nd ed, Hoboken, NJ.

Pirooznia M, Goes FS, Zandi PP. 2015. Whole-genome CNV analysis: advances in computational approaches. Front Genet 6:138.

Prober JM, Trainor GL, Dam RJ, Hobbs FW, Robertson CW, Zagursky RJ, Cocuzza AJ, Jensen MA, Baumeister K. 1987. A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides. Science 238:336-41.

Quick J, Loman NJ, Duraffour S, Simpson JT, Severi E, et al. 2016. Real-time, portable genome sequencing for Ebola surveillance. Nature 530:228-32.

Rand AC, Jain M, Eizenga JM, Musselman-Brown A, Olsen HE, Akeson M, Paten B. 2017. Mapping DNA methylation with high-throughput nanopore sequencing. Nat Methods 14:411-3.

Rhoads A, Au KF. 2015. PacBio sequencing and its applications. Gen Proteomics Bioinf 13:278-89.

Rothberg JM, Hinz W, Rearick TM, Schultz J, Mileski W, Davey M, Leamon JH, et al. 2011. An integrated semiconductor device enabling non-optical genome sequencing. Nature 475:348-52.

Ruparel H, Bi L, Li Z, Bai X, Kim DH, Turro NJ, Ju J. 2005. Design and synthesis of a 3'-O-allyl photocleavable fluorescent nucleotide as a reversible terminator for DNA sequencing by synthesis. Proc Natl Acad Sci USA 102:5932-7.

Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, Fiddes CA, Hutchison CA, Slocombe PM, Smith M. 1977a. Nucleotide sequence of bacteriophage uX174 DNA. Nature 265:687-95.

Sanger F, Coulson AR, Hong GF, Hill DF, Petersen GB. 1982. Nucleotide sequence of bacteriophage λ DNA. J Mol Biol 162:729-73.

Sanger F, Coulson AR. 1975. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. J Mol Biol 94:441-8.

Sanger F, Nicklen S, Coulson AR. 1977b. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA 74:5463-7.

Schadt EE, Banerjee O, Fang G, Feng Z, Wong WH, Zhang X, Kislyuk A, Clark TA, Luong K, Keren-Paz A, Chess A, Kumar V, Chen-Plotkin A, Sondheimer N, Korlach J, Kasarskis A. 2013. Modeling kinetic rate variation in third generation DNA sequencing data to detect putative modifications to DNA bases. Genome Res 23:129-41.

Seo TS, Bai X, Kim DH, Meng Q, Shi S, Ruparel H, Li Z, Turro NJ, Ju J. 2005. Four-color DNA sequencing by synthesis on a chip using photocleavable fluorescent nucleotides. Proc Natl Acad Sci USA 102:5926-31.

Smith LM, Fung S, Hunkapiller MW, Hunkapiller TJ, Hood LE. 1985. The synthesis of oligonucleotides containing an aliphatic amino group at the 50 terminus: synthesis of fluorescent DNA primers for use in DNA sequence analysis. Nucleic Acids Res 13:2399-412.

Smith LM, Sanders JZ, Kaiser RJ, Hughes P, Dodd C, Connell CR, Heiner C, Kent SB, Hood LE. 1986. Fluorescence detection in automated DNA sequence analysis. Nature 321:674-9.

Staden R. 1979. A strategy of DNA sequencing employing computer programs. Nucleic Acids Res 6:2601-10.

Tabor S, Richardson CC. 1987. DNA sequence analysis with a modified bacteriophage T7 DNA polymerase. Proc Natl Acad Sci USA 84:4767-71.

Tang H, Lyons E, Town CD. 2015. Optical mapping in plant comparative genomics. GigaScience 4:3.

Tawfik DS, Griffiths AD. 1998. Man-made cell-like compartments for molecular evolution. Nat Biotechnol 16:652-6.

Teague B, Waterman MS, Goldstein S, Potamousis K, Zhou S, Reslewic S, Sarkar D, Valouev A, Churas C, Kidd JM, Kohn S, Runnheim R, Lamers C, Forrest D, Newton MA, Eichler EE, Kent-First M, Surti U, Livny M, Schwartz DC. 2010. High-resolution human genome structure by single-molecule analysis. Proc Natl Acad Sci USA 107:10848-53.

Teh BT, Lim K, Yong CH, Ng CCY, Rao SR, Rajasegaran V, Lim WK, Ong CK, Chan K, Cheng VKY, Soh PS, Swarup S, Rozen SG, Nagarajan N, Tan P. 2017. The draft genome of tropical fruit durian (*Durio zibethinus*). Nat Genet 49:1633-41.

Thorisson GA, Smith AV, Krishnan L, Stein LD. 2005. The international HapMap project web site. Genome Res 15:1592-3.

Thu KL, Vucic EA, Kennett JY, Heryet C, Brown CJ, Lam WL, Wilson IM. 2009. Methylated DNA immunoprecipitation. J Vis Exp 23:935.

Veluchamy A, Rastogi A, Lin X, Lombard B, Murik O, Thomas Y, Dingli F, Rivarola M, Ott S, Liu X, Sun Y, Rabinowicz PD, McCarthy J, Allen AE, Loew D, Bowler C, Tirichine L. 2015. An integrative analysis of post-translational histone modifications in the marine diatom *Phaeodactylum tricornutum*. Genome Biol 16:102.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, et al. 2001. The sequence of the human genome. Science 291:1304-51.

Wallberg A, Bunikis I, Pettersson OV, Mosbech MB, Childers AK, Evans JD, Mikheyev AS, Robertson HM, Robinson GE, Webster MT. 2019. A hybrid de novo genome assembly of the honeybee, Apis mellifera, with chromosome-length scaffolds. BMC Genomics 20:275.

Wenger AM, Peluso P, Rowell WJ, Chang PC, Hall RJ, Concepcion GT, Ebler J, Fungtammasan A, Kolesnikov A, Olson ND, Töpfer A, Alonge M, Mahmoud M, Qian Y, Chin CS, Phillippy AM, Schatz MC, Myers G, DePristo MA, Ruan J, Marschall T, Sedlazeck FJ, Zook JM, Li H, Koren S, Carroll A, Rank DR, Hunkapiller MW. 2019. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nat Biotechnol 37:1155-62.

Wright B, Farquharson KA, McLennan EA, Belov K, Hogg CJ, Grueber CE. 2019. From reference genomes to population genomics: comparing three reference-aligned reduced-representation sequencing pipelines in two wildlife species. BMC Genomics 20:453.

Zhang J, Fang Y, Hou JY, Ren HJ, Jiang R, Roos P, Dovichi NJ. 1995. Use of non-cross-linked polyacrylamide for four-color DNA sequencing by capillary electrophoresis separation of fragments up to 640 bases in length in two hours. Anal Chem 67:4589-93.

CHAPTER 2

AN OPEN-SOURCE PROGRAM (HAPLO-ST) FOR WHOLE-GENOME SEQUENCE TYPING SHOWS EXTENSIVE DIVERSITY AMONG *LISTERIA MONOCYTOGENES* ISOLATES IN OUTDOOR ENVIRONMENTS AND POULTRY PROCESSING PLANTS[1]

[1] Swarnali Louha, Richard J. Meinersmann, Zaid Abdo, Mark E. Berrang, and Travis C. Glenn, Appl Environ Microbiol, 87(1), DOI:10.1128/AEM.02248-20, Reprinted here with permission of publisher.

2.1 ABSTRACT

A reliable and standardized classification of *Listeria monocytogenes* is important for accurate strain identification during outbreak investigations. Current whole-genome sequencing (WGS)-based approaches for strain characterization are either difficult to standardize, rendering them less suitable for data exchange, or are not freely available. Thus, we developed a portable and open-source tool, Haplo-ST, to improve standardization and provide maximum discriminatory potential to WGS data tied to a multi-locus sequence typing (MLST) framework. Haplo-ST performs whole-genome MLST (wgMLST) for *L. monocytogenes* while allowing for data exchangeability worldwide. This tool takes in (i) raw WGS reads as input, (ii) cleans the raw data according to user specified parameters, (iii) assembles genes across loci by mapping to genes from reference strains, and (iv) assigns allelic profiles to assembled genes and provides a wgMLST subtyping for each isolate. Data exchangeability relies on the tool assigning allelic profiles based on a centralized nomenclature defined by the widely-used BIGSdb-*Lm* database. Tests of Haplo-ST's performance with simulated reads from *L. monocytogenes* reference strains demonstrated high sensitivity (97.5%), and coverage depths ≥ 20× were found to be sufficient for wgMLST profiling. We then used Haplo-ST to characterize and differentiate between two groups of *L. monocytogenes* isolates derived from the natural environment and poultry processing plants. Phylogenetic reconstruction identified lineages within each group, and no lineage-specificity was observed with isolate phenotypes (transient vs. persistent) or origins. Genetic differentiation analyses between isolate groups identified 21 significantly differentiated loci, potentially enriched for adaptation and persistence of *L. monocytogenes* within poultry processing plants.

2.2 IMPORTANCE

We have developed an open-source tool (https://github.com/swarnalilouha/Haplo-ST) that provides allele-based subtyping of *L. monocytogenes* isolates at the whole genome level. Along with allelic profiles, this tool also generates allele sequences and identifies paralogs, which is useful for phylogenetic tree reconstruction and deciphering relationships between closely related isolates. More broadly, Haplo-ST is flexible and can be adapted to characterize the genome of any haploid organism simply by installing an organism-specific gene database. Haplo-ST also allows for scalable subtyping of isolates; fewer reference genes can be used for low resolution typing, whereas higher resolution can be achieved by increasing the number of genes used in the analysis. Our tool enabled clustering of *L. monocytogenes* isolates into lineages and detection of potential loci for adaptation and persistence in food processing environments. Findings from these analyses highlight the effectiveness of Haplo-ST in subtyping and evaluating relationships among isolates in studies of bacterial population genetics.

2.3 INTRODUCTION

Listeria monocytogenes is an opportunistic foodborne pathogen associated with significant public health concern worldwide, with an estimated 1600 illnesses and 260 deaths occurring annually (Bennion et al. 2008, Scallan et al. 2011) and an estimated annual economic burden of $2.8 billion in the United States (USDA ERS 2014). *L. monocytogenes* primarily causes the food borne illness listeriosis but may also cause septicemia, encephalitis, and meningitis in the immunocompromised, newborn, and elderly and severe complications in pregnancies leading to stillbirths and miscarriages (Den Bakker et al. 2008).

Listeriosis mainly occurs through the consumption of food such as meat, fish, and dairy products which become contaminated in food processing facilities during manufacturing, post-processing, or storage for extended periods of time before consumption (Painset et al. 2019). Within food processing facilities, *L. monocytogenes* can adapt to survive conditions used for food preservation and safety; it can replicate at low temperatures and under high-salt conditions and can withstand disinfectants and nitrate preservation methods. These, together with the ability to form biofilms on food contact surfaces, can facilitate prolonged persistence of *L. monocytogenes* in food facilities (Orsi et al. 2008, Carpentier and Cerf 2011, Hyden et al. 2016). Persistence may also arise from the survival of the bacteria in nooks not reached by regular cleaning and sanitation procedures. Often, this results in cross-contamination of the final product multiple times, which increases the risk of an outbreak. On the other hand, frequent introduction of *L. monocytogenes* from external sources may result in a high prevalence of transient strains within food facilities (Jagadeesan et al. 2019). Contaminating strains of *L. monocytogenes* are later released from food facilities into the natural environment via effluents (Berrang et al. 2005, Kuhn and Goebel 2007). Hence, food regulatory authorities frequently implement effective surveillance and control measures to discriminate between transient and persistent strains, decrease harborage, and prevent dissemination of *L. monocytogenes* (Jackson et al. 2016). Additionally, it is important to investigate the relatedness of strains of *L. monocytogenes* involved in a single contamination event for accurate source tracking. Such investigations can help optimize effective control measures to prevent recurrence of contamination in food processing facilities (Jagadeesan et al. 2019).

Molecular subtyping techniques have been traditionally used for strain discrimination and identification of degrees of genetic relatedness among isolates (Moorman et al. 2010). While many other subtyping methods (ribotyping, repetitive extragenic palindromic PCR [REP-PCR], and multilocus enzyme electrophoresis [MLEE]) have been used in the past, pulsed-field gel electrophoresis (PFGE) has been the “gold standard” subtyping tool for *L. monocytogenes* for many years (Swaminathan et al. 2001). Although PFGE has been extremely useful in outbreak investigations and source tracking of *L. monocytogenes* at food settings (Jagadeesan et al. 2019), it is time-consuming, labor-intensive, expensive, and difficult to standardize (Ruppitsch et al. 2015, Henri et al. 2016). Moreover, it provides little information on the genetic variation within or phylogenetic relationships among strains, limiting our overall understanding of evolutionarily important traits such as virulence. In contrast, sequence-based approaches are promising tools for strain typing and phylogeny assessment (Ragon et al. 2008). Multi-locus sequence typing (MLST) differentiates strains by detecting variation within the nucleotide sequences of seven housekeeping genes. Every isolate is defined by a sequence type (ST), which consists of a combination of seven allelic profiles. Groups of STs sharing a minimum of six identical alleles along with an ST acting as the ‘central genotype’ forms clonal complexes (CCs), which can be geographically and temporally widespread (Ragon et al. 2008). Conventional MLST has been used to describe the population structure of *L. monocytogenes*, and has shown that *L. monocytogenes* forms a structured population consisting of four divergent lineages (I-IV) (Ragon et al. 2008, Orsi et al. 2011). Lineage I strains are known to be highly clonal, indicating strong selection of genetic traits of fitness within the host, whereas Lineage II strains show higher rates of recombination than Lineage I, and this increased genome plasticity may help in adapting to diverse ecological niches (Meinersmann et al. 2004, Pirone-Davies et al. 2018). This is supported by the fact that Lineage I strains are predominantly linked to human clinical infection and animal listeriosis, whereas Lineage II strains are more commonly associated with food contamination and the environment. Lineage III and IV strains occur less frequently among humans and have been linked to animals (Dreyer et al. 2016).

The advent of next generation sequencing technologies has facilitated whole genome sequencing (WGS)-based subtyping at low costs and speeds exceeding that of traditional MLST. WGS enables easy availability of total bacterial genomes that allow strain discrimination at very high resolution. WGS also provides the ability to infer phylogenetic relationships among isolates, along with access to additional information, such as virulence and resistance markers (Painset et al. 2019). WGS-based subtyping has been used for the strain detection and surveillance of *L. monocytogenes* in different countries around the world (Jackson et al. 2016, Kvistholm Jensen et al. 2016, Kwong et al. 2016, Moura et al. 2017, Halbedel et al. 2018). WGS-based subtyping approaches are either based on single nucleotide polymorphisms (SNPs) (Jackson et al. 2016, Katz et al. 2017), or on gene-by-gene allelic profiling of a defined set of genes in the genome (Jagadeesan et al. 2019, Moura et al. 2017). Studies have shown that both SNP-based subtyping and whole-genome-based allelic profiling show similar discriminatory power and clustering among isolates (Jagadeesan et al. 2019, Henri et al. 2017). However, SNP-based approaches are dependent on the choice of the assembly pipeline and multiple closely related reference genomes which lack a global consensus, thus making standardization of SNP-based approaches among different laboratories difficult (Henri et al. 2017, Pearce et al. 2018). These limitations are overcome by gene-by-gene approaches (see Table S1 in the supplemental material), which are based on allelic variation of a predefined set of genes from either the core genome (core genome MLST [cgMLST]) or on a set of genes from both core and accessory genome (whole-genome MLST [wgMLST]). Several cgMLST schemes have been developed for subtyping *L. monocytogenes* (Ruppitsch et al. 2015, Pightling et al. 2015, Chen et al. 2016, Moura et al.

2016). These cgMLST schemes are different from each other with respect to the method employed, the diversity and number of isolates used in scheme development, and the number of loci used in each scheme. These differences between cgMLST schemes can impact communication on cluster detection between different laboratories, as knowledge on the type of core genome scheme, assembler, assembler version, and sequencing technology used for cluster detection becomes crucial (Pietzka et al. 2019). Furthermore, cgMLST finds differences only within the core genome of *L. monocytogenes*, which represents ~58% of the genome in terms of number of genes and ~54% in terms of the length of the genome (Jagadeesan et al. 2019). Though this level of differentiation may be sufficient for discriminating outbreak strains from epidemiologically unrelated strains, investigating persistence and source tracking of root-cause analysis requires increased discriminatory power beyond cgMLST (Jagadeesan et al. 2019). These problems can be addressed with a standardized wgMLST-based subtyping, which can profile allelic differences among *L. monocytogenes* strains on a genome-wide scale.

In this study, we present the Haploid Sequence-Typer (Haplo-ST), a tool that can perform wgMLST for *L. monocytogenes* while allowing for data exchangeability worldwide (Fig. 2.1 and Table 2.1). After developing Haplo-ST, we used it to characterize and differentiate between two groups of *L. monocytogenes* isolates: the first group was obtained from the natural environment, and the second group was obtained from poultry processing plants. Isolates obtained from the poultry processing plants contained both transient and persistent strains of *L. monocytogenes*. Previous research has shown that persistent strains have increased adhesion and biofilm formation capacity (Wang et al. 2015) and are genetically distinct from transient strains (Autio et al. 2003). However, larger-scale studies of the extent of genetic variation existing between persistent and transient strains are still needed.

This study aims (i) to develop Haplo-ST for performing wgMLST of *L. monocytogenes* isolates, (ii) to establish phylogenetic relationships within the two group of *L. monocytogenes* isolates obtained from the outdoor environment and poultry processing plants, (iii) to examine if there exists any lineage-specific association of isolates obtained from (a) different sites in the natural environment and (b) transient and persistent strains, and (iv) to analyze the extent of genetic variation between (a) isolates obtained from the natural environment and poultry processing plants and (b) transient and persistent strains of *L. monocytogenes*. We describe below how we achieved these aims.

2.4 RESULTS

2.4.1 Sensitivity of Haplo-ST

Allelic profiles derived from Haplo-ST for *L. monocytogenes* strains EGD-e and 4b F2365 were compared to allele profiles of 1826 loci in EGD-e and 1825 loci in 4b F2365 respectively. On average, 4.4% of genes had uncalled alleles; this may be due to the inability of short reads to assemble these genes completely. Amongst the loci that were assigned allele designations, reproducibility of allele calls with Haplo-ST was significant, yielding an average sensitivity of 97.5% over eight simulated datasets for coverage depths of ~ 80× (Phred quality score ≥ 20 for ≥ 90% bases in the retained reads).

2.4.2 Dependency of Haplo-ST on sequencing depth

The number of genes correctly profiled by Haplo-ST increased rapidly from a sequencing depth of 5× to 10×, then increased modestly from 10× to 20× and did not increase further beyond a depth of 20× (Fig. 2.2A). The number of genes assigned an erroneous allele ID (i.e., misassigned) and the number of genes missing an allele ID assignment (i.e., missing or uncalled alleles) decreased significantly up to a depth of 20×, improved slightly at 30×, and then remained stable at higher sequencing depths (Fig. 2.2B). The average number of genes partially assembled by YASRA and thus giving rise to uncalled alleles by BIGSdb-*Lm* remained similar over all sequencing depths. From these results, we conclude that sequencing depths ≥ 20× will perform well in Haplo-ST for wgMLST profiling of *L. monocytogenes* isolates.

2.4.3 wgMLST profiling of *L. monocytogenes* isolates

Haplo-ST generated a wgMLST profile of each *L. monocytogenes* isolate from WGS reads (see Data Set S4 in the supplemental material). A list of assembled gene sequences identified in each isolate were also provided by Haplo-ST (available at https://bit.ly/3e9KM6g).

2.4.4 Identification of paralogs

We used two different approaches to identify paralogous genes in our dataset. With our first approach, Haplo-ST generated a list of paralogous genes for each *L. monocytogenes* isolate while profiling isolates. Our second approach identified 133 paralogous genes (see Data Set S1 in the supplemental material) in BIGSdb-*Lm*. Comparison of the two approaches for paralog detection showed that BIGSdb-*Lm* correctly identifies all paralogous genes. However, in a few instances BIGSdb-*Lm* incorrectly identifies genes which are not paralogous to each other as ‘exact matches’ to each other (see Fig. S1 in the supplemental material). On further examination, we found that in such cases, two allele sequences partially matched across their lengths with a 100% identity (see examples in Fig. S1).

2.4.5 Population structure and phylogenetic relationships among *L. monocytogenes* isolates

Of the 171 *L. monocytogenes* isolates obtained from the Broad river watershed, 23 different CCs and 25 singleton STs (unassigned CCs) were identified (Data Set S4). Thirty-one novel STs were also identified in this group, revealing significant amount of diversity of *L. monocytogenes* strains. The distribution of the 23 different CCs in river water flowing through different land use areas is shown in Fig 2.3A. Of the 23 different CCs, 5 CCs (CC945, CC14, CC901, CC912, and CC910) were found to be the most abundant in this group. Across the population, some CCs were significantly enriched in water flowing through forests (CC14 and CC945), and others were more associated with pastures (CC901). The 171 isolates formed three distinct clusters in the phylogenetic tree (Fig. 2.4A), with each cluster containing a specific lineage (I, II and III) of *L. monocytogenes* strains. A majority of the isolates belonged to lineage II (68%), followed by isolates from lineage III (17%), and then lineage I (5%), and 15 isolates (9%) that could not be genotyped into lineages with lineage-specific probes clearly clustered in lineage II. A few isolates were distantly related to these clusters, and an assembly of the 16s rRNA sequence of these isolates showed that they belonged to non-pathogenic *Listeria* species, *L. seeligeri* (n=2) and *L. welshimeri* (n=1). Lineage I contained isolates from CC1, CC4, and three singletons (CC388, ST898 and a new ST), whereas most isolates in lineage III had novel STs, except for two isolates belonging to ST978. Lineage II was subdivided into seven clades that correspond mainly as follows: (1) CC940, CC950, CC912, ST941, ST936, and six singletons (ST914, ST949, ST985, ST990, ST913, and ST947); (2) CC945, CC935, ST956, ST951 and six singletons (ST944, CC838, ST948, ST939, ST909, and ST955); (3) CC926, CC920, and a singleton, CC831; (4) CC7, CC11, CC931, and seven singletons (CC177, CC918, CC14, CC906, ST934, ST789, and ST916); (5) ST390 and ST899; (6) CC14, CC901, CC570, and a singleton, ST954; and (7) CC321, CC910, and a singleton, ST929. Isolates from the three *L. monocytogenes* lineages were found to be randomly distributed across the four sampling locations (Fig. 2.4B). This was confirmed with Fisher's exact test (α = 0.05), which failed to show any lineage-specific association of isolates with sampling sites (*P* = 0.067). Because lineage III strains are mostly associated with animals (Dreyer et al. 2016), we hypothesized that the majority of lineage III strains would be obtained from agricultural/pastoral sites. However, in our data, most lineage III strains (62%) were obtained from forested areas.

The 162 isolates obtained from poultry processing plants contained 16 CCs and 1 singleton ST, ST1006 (Data Set S4). Nine isolates could not be assigned to any ST or CC, either due to new alleles identified in these isolates or due to incomplete MLST profiles. Six CCs (CC321, CC5, CC155, CC6, CC7, and CC9) accounted for 84% of the *L. monocytogenes* isolates (Fig 2.3B). Four CCs were abundant in the persistent strains: CC5, CC6, CC155, and CC321. The phylogenetic tree constructed from isolates obtained from poultry processing plants had two major clusters (Fig. 2.5A): one containing isolates belonging to lineage I (35%) and the other containing lineage II (59%) isolates. Twelve isolates could not be classified into lineages by genotyping with lineage-specific probes; of these twelve, 3 isolates clustered in lineage I and 4 isolates clustered in lineage II. The remaining 5 isolates were distantly related from the two major lineages in the tree and were identified as non-pathogenic species of *Listeria*, *L. innocua* (n=4) and *L. welshimeri* (n=1). Lineage I had two clades corresponding to (1) CC6 and three singletons (CC1, CC2, and CC4) and (2) CC5 and two singletons (CC288 and CC224). Lineage II was subdivided into four main clades corresponding to (1) CC321, (2) CC155, (3) CC9 and CC8, and (4) CC7 and two singletons (CC199 and ST1006). Persistent strains were more abundant (65%) than transient strains (35%), and correlation of *L. monocytogenes* lineages with transient versus persistent phenotypes was not significant (Fishers exact test, *P* = 0.86) (Fig. 2.5B).

2.4.6 Analysis of molecular variance

The wgMLST profiles were filtered for paralogous genes and assigned custom allele ID's for new alleles (see Data Set S5 in the supplemental material). The results from AMOVA showed that most genetic variation was contained within isolates obtained from the natural environment and poultry processing plants (91%), with only 9% attributed to variation between the two groups (Table 2.2). To detect loci with significant genetic variation between the two groups, we calculated population specific F_{ST} values for each locus separately with locus-by-locus AMOVA. We chose 111 loci (top 5% of F_{ST} distribution; $F_{ST} \geq 0.149$), with the highest F_{ST} values as loci having considerable genetic variation between isolates obtained from the natural environment and poultry processing plants (Fig. 2.6A; see Data Set S2 in the supplemental material). Additionally, results from AMOVA considering only isolates from the poultry processing plants suggested that majority of the genetic variance was within isolates (96.18%) and the remaining variation (3.18%) was between the transient and persistent groups of strains (Table 2.2). In this case, 102 loci (upper 5%; $F_{ST} \geq 0.782$) were identified as having the most divergence between the transient and persistent strains (Fig. 2.6B; see Data Set S3 in the supplemental material). A set of 21 loci were common among the loci with highest F_{ST} values in both levels of AMOVA (i.e., 111 loci in natural environment versus poultry processing plants and 102 loci in transient versus persistent groups) and might play a role in the adaptation and persistence of *L. monocytogenes* in poultry processing environments (Table 2.3).

2.5 DISCUSSION

Molecular characterization of *Listeria monocytogenes* is important for outbreak detection, surveillance, and epidemiological studies and in the development of effective control strategies for listeriosis. We have developed a freely available and portable tool, Haplo-ST, that can be used for wgMLST profiling of *L. monocytogenes* from WGS data (Fig 2.1). In contrast to the commercial genome-wide MLST developed by BioNumerics® (Applied Maths NV, Belgium) and being used by the US CDC and PulseNet International, Haplo-ST is open-source (Table 2.1). Our tool uses the centralized nomenclature of *L. monocytogenes* genotypes publicly accessible in the BIGSdb-*Lm* database and the BIGSdb software for calling alleles, which facilitates sharing and comparing data between public health laboratories worldwide. We have shown that the reproducibility of allele calls by Haplo-ST has high sensitivity (error rate ~ 2.5%), and sequencing depths of ~20× are sufficient for assembling alleles (Fig 2.2). Because our genotyping technique assembles alleles directly from WGS data by mapping to corresponding reference genes before allele typing, it is computationally faster and less error-prone than other subtyping techniques that require *de novo* assembly of genomes prior to allele identification and subtyping (Ruppitsch et al. 2015, Moura et al. 2016). This property also allows for the scalable characterization of isolates based on the needs of the researcher, as some questions require more discrimination among isolates than others. For example, lower resolution is required for assignment of isolates to a specific lineage or clonal complex, whereas higher levels of discrimination are needed for outbreak detection and investigation of within-patient variations (Maiden et al. 2013). In this regard, Haplo-ST is flexible because it can be used with custom sets of fewer reference genes for low resolution typing, whereas higher resolution can be achieved by increasing the number of reference genes used in the analysis. The time required for low resolution typing is low and increases with the increase in typing resolution. For example, on a system with a quad-core processor running at 3.6 GHz and 50 GB of RAM, the time taken for subtyping 100, 500 and 1000 loci were 1.4, 6.2, and 12.8 h, respectively.

The motivation to develop Haplo-ST was to design a platform that can harness the full power of Illumina sequencing for characterizing *L. monocytogenes* isolates, thereby subtyping them at the highest possible level of resolution, which can be used for discriminating between closely related isolates that have diversified over a short timeframe. This is highly relevant during outbreak investigations and for tracking the origin of contamination, precise assessment of divergence dates, and forming hypotheses on the mechanisms of segregation of isolates. This discriminatory power of wgMLST is not achieved with cgMLST because it only assesses differences in the core genome and has been shown to provide fewer allelic differences in comparison to wgMLST (Jagadeesan et al. 2019). Furthermore, cgMLST schemes are mostly composed of slowly evolving genes. Previous studies on *L. monocytogenes* genomes have estimated the evolution rate of cgMLST types to be around 0.2 alleles per year, indicating that cgMLST-based typing is insufficient for discriminating isolates which have diverged over short timeframes (Moura et al. 2016). However, use of a well-defined set of species-wide conserved genes makes cgMLST more stable and suitable for robust comparisons of distantly related isolates. Typically, cgMLST is sufficient for routine epidemiological surveillance, such as identification of clonal groups and discrimination of outbreak strains from epidemiologically unrelated strains. Haplo-ST can perform both core-genome and whole-genome MLST because its database incorporates genes in the core-genome (the *L. monocytogenes* cgMLST scheme developed by Institut Pasteur) together with accessory genes in the pan-genome of *L. monocytogenes*. Additionally, it can be used for inferring biological properties, such as virulence, antibiotic-resistance, and stress tolerance, and phenotypic predictions like serotypes by profiling genes linked to these properties. The wgMLST currently provided with Haplo-ST can also be expanded to include genotypic variation in future *L. monocytogenes* isolates by updating the locally installed BIGSdb-*Lm* database housed within this platform. This can include multi-copy and accessory genes, which may arise through recombination and whose detection may become important for pathogen surveillance.

Unlike SNP-based genotyping which uses individual SNPs as units of comparison, cg- or wgMLST counts different types of variants within one coding region as a single allelic change. This concept covers the conflicting signals of horizontal and vertical transfer of genetic material as a single evolutionary event and classifies WGS data as a set of allele identifiers, thereby enabling easy storage of a stable nomenclature within a database and making comparisons of wgMLST profiles faster. Nonetheless, this also leads to a loss of resolution as it obscures the extent of dissimilarity between non-identical alleles. Thus, the technical performance of wgMLST, along with its amenability to standardization, is accompanied by a loss in specificity, as minimum spanning trees constructed using sequence types are fully connected, failing to effectively split isolate populations into clonal complexes (Feil et al. 2004). This becomes problematic as allele-based subtyping alone does not provide sufficient information for delineating outbreaks; it is therefore critical to complement it with whole-genome-based phylogenetic clustering for accessing relationships between isolates (Chen et al. 2017). Recent studies have shown that although wgMLST-based dendrograms are comparable to SNP-based phylogenies in identifying clades of closely related isolates with a recent common ancestor, they differ from each other with respect to the placement of isolates within clonal groups where branches in SNP-based phylogenies are not supported by greater than 90% bootstrap support

(Jagadeesan et al. 2019). This emphasizes the importance of constructing phylogenies with confidence measures such as bootstrap support, which is unfortunately not feasible with wgMLST-based dendrograms. Haplo-ST has the advantage of not only providing wgMLST profiles, but also provides corresponding allele sequences assembled for each isolate. While allelic profiles can be used for constructing dendrograms from allelic similarity-type matrices, allele sequences can be concatenated and used for constructing cg- or wgMLST-based phylogenies using a variety of models of molecular evolution and obtaining bootstrap support values. Moreover, our tool can detect paralogous genes, which when ignored can lead to the construction of biased phylogenies. Thus, analysis provided by Haplo-ST, when combined with detailed epidemiological evidence, isolate metadata, and appropriate interpretation, allows for routine surveillance of *L. monocytogenes*, accurate source-tracking of contaminating strains, and elucidation of transmission pathways and ultimately helps in devising better intervention strategies in food safety monitoring programs.

Our approach was evaluated for its usability in characterizing and determining relatedness within two groups of *L. monocytogenes* isolates: one group representing isolates present in the natural environment and the other from poultry further processing facilities. This enabled us to decipher the phylogenetic relatedness of *L. monocytogenes* isolates, which shows clear delineation between lineages in both isolate groups (Fig 2.4 and 2.5). A majority of isolates in the natural environment and food facilities belonged to lineage II, which is consistent with previous studies (Dreyer et al. 2016). Furthermore, the lineage of 11% isolates could not be identified with lineage-specific probes (Data Set S4). All of these were identified using our methods, including 2% that belonged to other species (Fig 2.4 and 2.5). Moreover, we did not find significant differences in the distribution of isolates belonging to different lineages in terms of their phenotypes (persistent/transient) and origin (sampling sites). However, it is curious that no lineage III isolates were found in the processing plant samples, although they made up 17% of isolates obtained from the natural environment. Distribution of CCs and STs between the groups of isolates showed that 5 CCs (CC1, CC4, CC7, CC11, and CC321) were common across both isolate groups (Fig 2.3).

Listeria monocytogenes is a foodborne pathogen that is ubiquitous in the natural environment. Its ability to colonize and persist in food processing environments increases the risk of contaminating ready-to-eat (RTE) food, often leading to outbreaks of listeriosis. Hence, understanding the genetic determinants associated with its adaptation and persistence in food processing plants can indicate specific traits selected in the processing plant environment and the genetic and physiological factors responsible for the persistent phenotype. This is of paramount importance for developing targeted intervention strategies in the food industry, and the typing of *L. monocytogenes* plays a crucial role in such investigations.

We used Haplo-ST to type and identify loci with significant genetic variation between isolates obtained from the natural environment and poultry processing facilities. Our analysis revealed 111 significantly differentiated loci (Fig 2.6A; Data Set S2) which may be involved in helping *L. monocytogenes* to adapt to high stress conditions within food processing environments, thereby increasing its risk of contaminating food. Unlike transient strains, which are frequently introduced into food facilities from the natural environment and easily removed with regular sanitation shifts, persistent strains have been reported to have enhanced capacity to adapt and survive in food production chains and are difficult to eradicate. Thus, we also used our tool to characterize and detect loci with high genomic differentiation between transient and persistent strains. We obtained 102 highly differentiated loci potentially enriched for the 'persistent' phenotype (Fig 2.6B; Data Set S3). Of these, 21 loci were common with the 111 loci we previously identified as potentially contributing towards adaptation in food processing facilities (Table 2.3). These loci were related to metabolism (*lmo0875, lmo2650, lmo1336, lmo1817, lmo1464,* and *lmo2640*), transport (*lmo0875, lmo2650, lmo1210, lmo2383, lmo1960,* and *lmo1205*), tRNA and ribosome biogenesis (*lmo1949, lmo2078,* and *lmo1294*), biosynthesis of secondary metabolites (*lmo1294* and *lmo2640*), translation (*lmo2548* and *lmo2073*), and oxidative stress (*lmo0964*). We also found that out of the 102 loci differentiated for persistence, three genes (*lmo1699, lmo0692,* and *lmo2020*) were found to be associated with chemotaxis, a process that plays a role in niche localization (Casey et al. 2014). Several studies have shown the presence of a five-gene stress survival islet, SSI-1, to contribute to the growth of *L. monocytogenes* under suboptimal conditions, like low pH and high salt concentrations (Ryan et al. 2010, Gómez et al. 2014). Our analyses found SSI-1 in a higher fraction of isolates (93%) from processing plants compared to the natural environment (17%). Other studies report resistance to quaternary ammonium compounds, like benzalkonium chloride (BC), in persistent strains (Cherifi et al. 2018). BC is commonly used as an agri-food sanitizer, and resistance to it is provided by the gene cassette *bcrABC*, in which *bcrAB* codes for the small multidrug resistance protein family transportera and *bcrC* codes for a transcriptional factor. Our subtyping results are in agreement with this; *bcrABC* was present in 72% of the isolates obtained from the effluents, but absent in isolates obtained from the natural environment. Among isolates collected from effluents, *bcrABC* was associated with a higher proportion of persistent strains (54%) when compared to transient strains (18%).

Our approach does, however, have a few limitations. Although the locally installed database within our platform is expandable to accommodate future genetic diversity in *L. monocytogenes*, it requires frequent manual upgrades as new alleles and genes become available. With the recent accessibility of BIGSdb-*Lm* at Pasteur Institut through RESTful API, this drawback can be resolved by making minor modifications to our pipeline which will allow the tool to interrogate the server at Pasteur Institut directly instead of calling alleles locally. Secondly, our approach is gene-centric and characterizes differences only in protein-coding genes; therefore, genetic variation in other genomic regions like pseudogenes and intergenic regions are not accounted for. Additionally, the use of short reads may produce faulty assemblies of accessory genes and repeat regions. With the decreasing costs and increased popularity of third-generation sequencing instruments, these limitations can be overcome with development of appropriate sequence assembly algorithms. Thus, the power of fully assembled genomes remain yet to be exploited. Nevertheless, the current wgMLST approach will be stable over time as new genes are added and maintain backwards compatibility with classical seven-gene MLST schemes.

The greatest advantage of Haplo-ST is that this platform is flexible and not limited to profiling of *Listeria monocytogenes* alone. It can be adapted to provide molecular characterization for any haploid organism, with the installation of an organism-specific gene database with associated allelic nomenclature, along with minor changes to the script that automates the pipeline. Furthermore, users are not limited to using publicly available gene databases because BIGSdb can accommodate any custom user-provided database.

2.6 MATERIALS AND METHODS

2.6.1 Development of Haplo-ST for wgMLST profiling of *L. monocytogenes* strains

We developed Haplo-ST to analyze wgMLST for *L. monocytogenes* (Fig. 2.1). This tool takes in raw WGS reads for each *L. monocytogenes* isolate and uses the FASTX-Toolkit v0.0.14 (Hannon

2010) to clean them according to user-specified parameters. It then uses YASRA v2.33, (available at https://github.com/aakrosh/YASRA) to assemble genes across loci by mapping to reference genes. We selected YASRA for assembling genes because YASRA is a comparative assembler which uses a template to guide the assembly of a closely related target sequence and can accommodate high rates of polymorphism between the template and target (Ratan 2009). Hence, this assembler can be used to assemble an allelic variant of a gene by mapping to a reference sequence, even when the target allele has diverged considerably from the reference gene sequence. Next, a local installation of the BIGSdb-*Lm* database (available at http://bigsdb.pasteur.fr/listeria, Jolley and Maiden 2010) is used by Haplo-ST to assign allelic profiles to the genes assembled with YASRA, thus generating a wgMLST profile for each isolate. The BIGSdb-*Lm* database contains allelic profiles of 2554 *L. monocytogenes* genes obtained from BIGSdb-*Lm* as of 2 June 2017. This pipeline has been automated with a Perl script and made portable by installation of all software dependencies along with a local installation of the BIGSdb-*Lm* database within a Linux Virtual Machine (VM). In addition to generating wgMLST profiles, Haplo-ST also outputs the list of gene sequences assembled for each isolate. Because BIGSdb-*Lm* can identify all paralogs associated with a query gene sequence as 'exact matches', our tool has also been automated to output a list of paralogs identified for each isolate.

2.6.2 Sensitivity of Haplo-ST

ART v2.5.8 (Huang et al. 2012) was used to simulate WGS reads for two reference genomes of *L. monocytogenes*, EGD-e (NCBI accession number NC_003210.1) and strain 4b F2365 (NCBI accession number NC_002973.6). The simulated WGS reads were of two different lengths (150 bp and 250 bp), and different qualities (one set of reads with high quality throughout the read length and the other with degrading quality over the length of the read). In total, 8 sets of simulated WGS reads were processed through Haplo-ST to generate 8 wgMLST profiles. Four of these wgMLST profiles were obtained from simulated reads generated from the *L. monocytogenes* EGD-e reference genome. Each of these 4 profiles were compared to the allelic profiles of annotated genes in EGD-e. The other four wgMLST profiles were obtained from reads derived from the strain 4b F2365 reference genome. These were compared to the allelic profiles of annotated genes in F2365. For each comparison, we calculated the percentage of genes correctly typed by Haplo-ST. Finally, we calculated the average sensitivity over eight comparisons.

2.6.3 Dependency of Haplo-ST on sequencing depth

To determine the levels of genome sequence coverage necessary for efficient whole-genome sequence typing, synthetic reads were simulated from the *L. monocytogenes* EGD-e reference genome with ART v2.5.8 for different sequencing depths ranging from 5× - 120× and typed with Haplo-ST (performed in triplicate). For each sequencing depth, the allelic profiles typed by our tool were compared to allelic profiles of annotated genes from the *L. monocytogenes* EGD-e reference genome. Finally, for each comparison, we calculated: (i) the number of genes correctly typed, (ii) the number of genes assigned an erroneous allele ID, (iii) the number of genes partially assembled, and (iv) the number of genes missing an allele ID assignment by Haplo-ST.

2.6.4 Analysis of *L. monocytogenes* strains collected from the natural environment and poultry processing plants

2.6.4.1 *Listeria monocytogenes* Isolate collection, DNA extraction and sequencing

L. monocytogenes isolates obtained from the natural environment were cultured from water and sediment samples collected at 16 locations in the South Fork Broad River watershed, located in Northeast Georgia (Bradshaw et al. 2016). Sampling locations were selected based on predominant land use by the National Land Cover Database and on-the-ground surveys. Samples were collected from 6 sites designated as agricultural/pastoral, 7 sites as forested, 2 sites as impacted by water pollution control plants (WPCP), and 1 site classified as mixed-use. *L. monocytogenes* isolates obtained from poultry processing plants were sampled from different locations within the poultry processing plants at different time periods (Berrang et al. 2005, Berrang et al. 2010). Some of these isolates were repeatedly isolated from multiple sites in the plants over an extended period of time and were designated as 'persistent' types (based on actA-sequence subtyping); other isolates sporadically isolated from the food processing facilities were classified as 'transient' strains. Each colony of *L. monocytogenes* isolate cultured from the samples was inoculated into 5ml of tryptic soy broth and grown overnight at 35 °C. DNA was extracted using the UltraClean® Microbial DNeasy Kit (Qiagen, Venlo, The Netherlands) according to manufacturer's instructions. Sequencing libraries were prepared using the Nextera XT DNA Library Preparation Kit (Illumina, San Diego, USA). Genomic DNA of each isolate was sequenced using the Illumina MiSeq platform to obtain paired-end 150- or 250-bp reads. This effort yielded WGS data for a total of 171 *L. monocytogenes* isolates obtained from the natural environment (NCBI BioProject Accession: PRJNA605751) and 162 isolates obtained from poultry processing plants (NCBI BioProject Accession: PRJNA606479). Of the 162 isolates obtained from poultry processing plants, 57 isolates were transient and 105 isolates were persistent types (Data Set S4). These were then processed using Haplo-ST.

2.6.4.2 wgMLST profiling of *L. monocytogenes* isolates with Haplo-ST

WGS data for *L. monocytogenes* isolates was first checked for quality with FastQC v0.11.4 (Andrews 2010). The raw data was then cleaned with the FASTX-Toolkit v0.0.14 incorporated within Haplo-ST. User-specified parameters were used to perform three successive cleaning steps with FASTA/Q Trimmer, FASTQ Quality Trimmer, and FASTQ Quality Filter tools of the FASTX-Toolkit. Reads were trimmed to remove all bases with a Phred quality score of < 20 from both ends and filtered such that 90% of bases in the clean reads had a quality of at least 20. After trimming and filtering, all remaining reads with lengths of < 50 bp were filtered out. Next, the cleaned reads were assembled into gene sequences by mapping to reference genes with YASRA. While assembling genes across loci, all assemblies having a length of less than 89% of the length of the corresponding reference gene were removed. This is because our examination of the lengths of all 2554 genes and their respective alleles in the BIGSdb-*Lm* database revealed that alleles of a gene can have different lengths, which ranges from 0.89 - 1.09 times the length of the reference gene. This 'length criteria' for filtering assembled genes has been provided as a user-specified parameter in the Perl script that automates Haplo-ST. The value for this parameter can be adjusted if the BIGSdb-*Lm* database is updated to include more genes or alleles, or if only a subset of genes is used for allelic profiling. Finally, assembled genes were assigned allele ID's with BIGSdb-*Lm*, and wgMLST profiles were generated for each isolate. Each isolate was assigned an MLST sequence type (ST) and clonal complex (CC) in accordance with BIGSdb-*Lm*.

2.6.4.3 Identification of paralogous genes

We identified paralogous genes in our dataset using two approaches. In the first approach, Haplo-ST uses BIGSdb-*Lm*'s ability to identify paralogs and outputs a list of paralogs for each isolate. To verify that all paralogs were correctly identified with BIGSdb-*Lm*, we used a second approach to detect paralogs present within the BIGSdb-*Lm* database. First a local BLAST database was created with all 2554 genes and their corresponding alleles present in the BIGSdb-*Lm* database using BLAST+ v2.2.29. Next, BLAST searches of all genes and their respective alleles were made against the local BLAST database. Custom Perl scripts were used to identify genes having an exact sequence match to another gene in the database, and all such matches were listed as paralogs.

2.6.4.4 Construction of phylogenetic trees and evaluation of lineage-specific association

The list of genes assembled for each isolate with Haplo-ST were filtered to remove paralogous genes. The final filtered assemblies for each group of isolates (the first group obtained from the natural environment and the second group obtained from poultry processing plants) were used to create concatenated multiple sequence alignments (MSA) with Phyluce v1.5.0 (Faircloth 2016). Several scripts were used to create MSA's for each isolate group. First, a custom Perl script was used to convert the assembled gene sequences into a format suitable for use with Phyluce. Second, the 'phyluce_align_seqcap_align' script was used to align genes across loci for all isolates within a group and the alignment was trimmed for ragged edges. The summary statistics of alignments for both isolate groups were checked with the script 'phyluce_align_get_align_summary_data' and cleaned for locus names with 'phyluce_align_remove_locus_name_from_nexus_lines'. The dataset for each isolate group was then culled to reach a 95% level of completeness with 'phyluce_align_get_only_loci_with_min_taxa'. The 95% complete data matrix was converted into phylip files with 'phyluce_align_format_nexus_files_for_raxml' and phylogenetic trees were constructed with FastME v2.1.5 (Lefort et al. 2015). The substitution model used by FastME was 'p-distance', and the BioNJ algorithm was used to compute a tree from the distance matrix. A total of 500 bootstrap replicates were computed to provide support to the internal branches of each of the phylogenies.

Listeria monocytogenes isolates were classified into lineages (I to IV) based on a targeted multilocus genotyping approach (TMLGT) in which six genomic regions were coamplified in a multiplexed PCR and used as templates for allele-specific primer extension using lineage-specific probes (Ward et al. 2010). Lineage-specific correlation between groups of isolates was tested with Fisher's exact test at $P = 0.05$.

Phylogenetic trees were visualized and annotated with iTOL v3 (Letunic and Bork 2016). For better visualization, all phylogenetic trees were converted to circular format and lineage classification for isolates was displayed by coloring internal branches. The annotations for the source and type of isolates were displayed in outer external rings.

2.6.4.5 Analysis of genetic variation

To obtain measures of genetic differentiation, we used the wgMLST profiles from Haplo-ST and performed Analysis of Molecular Variance (AMOVA) in Arlequin v3.5.2 (Excoffier and Lischer 2010). First, paralogous loci were removed from the raw wgMLST profiles. Next, new alleles not defined in the BIGSdb-*Lm* database and reported as 'closest matches' to existing alleles in the wgMLST profiles were assigned custom allele ID's with in-house Python scripts. Finally, AMOVA was separately performed at two levels: (i) among groups of isolates obtained from the natural environment and poultry processing plants and (ii) among groups of transient and persistent strains obtained from the poultry processing plants. For each level of analysis, loci with < 10% missing data in the wgMLST profiles were used. Fifty thousand permutations were used to determine significance of variance components. In addition to the standard AMOVA, which calculates the global F_{ST} for all loci within a group of isolates, we also performed a locus-by-locus AMOVA, which computes F_{ST} indices for each locus separately, for both levels of analysis. The upper 5% of the distribution of F_{ST} values was chosen as the threshold for loci with significant genetic diversity.

2.7 ACKNOWLEDGEMENTS

This research was supported by funding from USDA Agricultural Research Service Project Number 6040-32000-009-00-D. We thank USDA and FSIS for providing us with *Listeria monocytogenes* whole-genome sequencing samples for our work. We also thank Yecheng Huang for assistance with a local installation of the BIGSdb database. The high-performance computing cluster at Georgia Advanced Computing Resource Center (GACRC) at the University of Georgia provided computational infrastructure and technical support throughout the work.

2.8 REFERENCES

Andrews S. 2010. FastQC: a quality control tool for high throughput sequence data. Repository http://www.bioinformatics.babraham.ac.uk/projects/fastqc

Autio T, Keto-Timonen R, Lundén J, Björkroth J, Korkeala H. 2003. Characterization of persistent and sporadic *Listeria monocytogenes* strains by pulsed-field electrophoresis (PFGE) and amplified fragment length polymorphism (ALFP). Syst Appl Microbiol 26:539-45.

Bennion JR, Sorvillo F, Wise ME, Krishna S, Mascola L. 2008. Decreasing listeriosis mortality in the United States, 1990-2005. Clin Infect Dis 47:867-74.

Berrang ME, Meinermann RJ, Frank JF, Ladely SR. 2010. Colonization of a newly constructed commercial chicken further processing plant with *Listeria monocytogenes*. J Food Prot 73:286-291.

Berrang ME, Meinersmann RJ, Frank JF, Smith DP, Genzlinger LL. 2005. Distribution of *Listeria monocytogenes* subtypes within a poultry further processing plant. J Food Prot 68:980-985.

Bradshaw JK, Snyder BJ, Oladeinde A, Spidle D, Berrang ME, Meinersmann RJ, Oakley B, Sidle RC, Sullivan K, Molina M. 2016. Characterizing relationships among fecal indicator bacteria, microbial source tracking markers, and associated waterborne pathogen occurrence in stream water and sediments in a mixed land use watershed. Water Res 101:498-509.

Carpentier B, Cerf O. 2011. Review–persistence of *Listeria monocytogenes* in food industry equipment and premises. Int J Food Microbiol 145:1-8.

Casey A, Fox EM, Schmitz-Esser S, Coffey A, McAuliffe O, Jordan K. 2014. Transcriptome analysis of *Listeria monocytogenes* exposed to biocide stress reveals a multi-system response involving cell wall synthesis, sugar uptake, and motility. Front Microbiol 5:68.

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

Chen Y, Luo Y, Carleton H, Timme R, Melka D, Muruvanda T, Wang C, Kastanis G, Katz LS, Turner L, Fritzinger A, Moore T, Stones R, Blankenship J, Salter M, Parish M, Hammack TS, Evans PS, Tarr CL, Allard MW, Strain EA, Brown EW. 2017. Whole Genome and Core Genome Multilocus Sequence Typing and Single Nucleotide Polymorphism Analyses of *Listeria monocytogenes* Isolates Associated with an Outbreak Linked to Cheese, United States, 2013. Appl Environ Microbiol 83:e00633-17.

Cherifi T, Carrillo C, Lambert D, Miniaï I, Quessy S, Larivière-Gauthier G, Blais B, Fravalo P. 2018. Genomic Characterization of *Listeria Monocytogenes* Isolates Reveals That Their Persistence in a Pig Slaughterhouse Is Linked to the Presence of Benzalkonium Chloride Resistance Genes. BMC Microbiol 18:220.

Den Bakker HC, Didelot X, Fortes ED, Nightingale KK, Wiedmann M. 2008. Lineage specific recombination rates and microevolution in *Listeria monocytogenes*. BMC Evol Biol 8:277.

Dreyer M, Aguilar-Bultet L, Rupp S, Guldimann C, Stephan R, Schock A, Otter A, Schüpbach G, Brisse S, Lecuit M, Frey J, Oevermann A. 2016. *Listeria monocytogenes* sequence type 1 is predominant in ruminant rhombencephalitis. Sci Rep 6:36419.

Excoffier L, Lischer HEL. 2010. Arlequin suite ver 3.5: A new series of programs to perform population genetics analyses under Linux and Windows. Mol Ecol Resour 10:564-567.

Faircloth BC. 2016. PHYLUCE is a software package for the analysis of conserved genomic loci. Bioinformatics 32:786-8.

Feil EJ, Li BC, Aanensen DM, Hanage WP, Spratt BG. 2004. eBURST: inferring patterns of evolutionary descent among clusters of related bacterial genotypes from multilocus sequence typing data. J Bacteriol 186:1518-30.

Gómez D, Azón E, Marco N, Carramiñana JJ, Rota C, Ariño A, Yangüela J. 2014. Antimicrobial Resistance of *Listeria Monocytogenes* and *Listeria Innocua* from Meat Products and Meat-Processing Environment. Food Microbiol 42:61-5.

Halbedel S, Prager R, Fuchs S, Trost E, Werner G, Flieger A. Whole-Genome Sequencing of Recent *Listeria monocytogenes* Isolates from Germany Reveals Population Structure and Disease Clusters. 2018. J Clin Microbiol 56:e00119-18.

Hannon GJ. 2010. FASTX-Toolkit, FASTQ/A short-reads pre-processing tools. Repository http://hannonlab.cshl.edu/fastx_toolkit

Henri C, Félix B, Guillier L, Leekitcharoenphon P, Michelon D, Mariet JF, Aarestrup FM, Mistou MY, Hendriksen RS, Roussel S. 2016. Population Genetic Structure of *Listeria monocytogenes* Strains as Determined by Pulsed-Field GelElectrophoresis and Multilocus Sequence Typing. Appl Environ Microbiol. 82:5720-8.

Henri C, Leekitcharoenphon P, Carleton HA, Radomski N, Kaas RS, Mariet JF, Felten A, Aarestrup FM, Gerner Smidt P, Roussel S, Guillier L, Mistou MY, Hendriksen RS. 2017. An Assessment of Different Genomic Approaches for Inferring Phylogeny of *Listeria monocytogenes*. Front Microbiol 8:2351.

Huang W, Li L, Myers JR, Marth GT. 2012. ART: A Next-Generation Sequencing Read Simulator. Bioinformatics 28:593-4.

Hyden P, Pietzka A, Lennkh A, Murer A, Springer B, Blaschitz M, Indra A, Huhulescu S, Allerberger F, Ruppitsch W, Sensen CW. 2016. Whole genome sequence-based serogrouping of *Listeria monocytogenes* isolates. J Biotechnol 235:181-6.

Jackson BR, Tarr C, Strain E, Jackson KA, Conrad A, Carleton H, Katz LS, Stroika S, Gould LH, Mody RK, Silk BJ, Beal J, Chen Y, Timme R, Doyle M, Fields A, Wise M, Tillman G, Defibaugh-Chavez S, Kucerova Z, Sabol A, Roache K, Trees E, Simmons M, Wasilenko J, Kubota K, Pouseele H, Klimke W, Besser J, Brown E, Allard M, Gerner-Smidt P. 2016. Implementation of Nationwide Real-time Whole-genome Sequencing to Enhance Listeriosis Outbreak Detection and Investigation. Clin Infect Dis 63:380-386.

Jagadeesan B, Baert L, Wiedmann M, Orsi RH. 2019. Comparative Analysis of Tools and Approaches for Source Tracking *Listeria monocytogenes* in a Food Facility Using Whole-Genome Sequence Data. Front Microbiol 10:947.

Jolley KA, and Maiden MC. 2010. BIGSdb:scalable analysis of bacterial genome variation at the population level. BMC Bioinform 11:595.

Katz LS, Griswold T, Williams-Newkirk AJ, Wagner D, Petkau A, Sieffert C, Domselaar GV, Deng X, Carleton HA. 2017. A comparative analysis of the Lyve-SET phylogenomics pipeline for genomic epidemiology of foodborne pathogens. Front Microbiol 8:375.

Kuhn M, Goebel W. 2007. Molecular virulence determinants of *Listeria monocytogenes*, p 111-155. *In* Ryser ET, Marth EH (ed), *Listeria*, listeriosis and food safety, 3rd ed, CRC Press Taylor and Francis Group, Boca Raton, FL.

Kvistholm Jensen A, Nielsen EM, Björkman JT, Jensen T, Müller L, Persson S, Bjerager G, Perge A, Krause TG, Kiil K, Sørensen G, Andersen JK, Mølbak K, Ethelberg S. 2016. Whole-genome Sequencing Used to Investigate a Nationwide Outbreak of Listeriosis Caused by Ready-to-eat Delicatessen Meat, Denmark, 2014. Clin Infect Dis 63:64-70.

Kwong JC, Mercoulia K, Tomita T, Easton M, Li HY, Bulach DM, Stinear TP, Seemann T, Howden BP. 2016. Prospective Whole-Genome Sequencing Enhances National Surveillance of *Listeria monocytogenes*. J Clin Microbiol 54:333-342.

Lefort V, Desper R, Gascuel O. 2015. FastME 2.0: A Comprehensive, Accurate, and Fast Distance-Based Phylogeny Inference Program. Mol Biol Evol 32:2798-800.

Letunic I, Bork P. 2016. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic Acids Res 44:W242-5.

Maiden MC, Jansen van Rensburg MJ, Bray JE, Earle SG, Ford SA, Jolley KA, McCarthy ND. 2013. MLST revisited: the gene-by-gene approach to bacterial genomics. Nat Rev Microbiol 11:728-36.

Meinersmann RJ, Phillips RW, Wiedmann M, Berrang ME. 2004. Multilocus Sequence Typing of *Listeria Monocytogenes* by Use of Hypervariable Genes Reveals Clonal and Recombination Histories of Three Lineages. Appl Environ Microbiol 70:2193-203.

Moorman M, Pruett P, Weidman M. 2010. Value and Methods for Molecular Subtyping of Bacteria, p 157-175. In Kornacki JL (ed), Principles of Microbiological Troubleshooting in the Industrial Food Processing Environment, 1st ed, Springer Science Business Media, New York, NY.

Moura A, Criscuolo A, Pouseele H, Maury MM, Leclercq A, Tarr C, Björkman JT, Dallman T, Reimer A, Enouf V, Larsonneur E, Carleton H, Bracq-Dieye H, Katz LS, Jones L, Touchon M, Tourdjman M, Walker M, Stroika S, Cantinelli T, Chenal-Francisque V, Kucerova Z, Rocha EPC, Nadon C, Grant K, Nielsen EM, Pot B, Gerner-Smidt P, Lecuit M, Brisse S. 2016. Whole genome-based population biology and epidemiological surveillance of *Listeria monocytogenes*. Nat Microbiol 2:16185.

Moura A, Tourdjman M, Leclercq A, Hamelin E, Laurent E, Fredriksen N, Van Cauteren D, Bracq-Dieye H, Thouvenot P, Vales G, Tessaud-Rita N, Maury MM, Alexandru A, Criscuolo A, Quevillon E, Donguy MP, Enouf V, de Valk H, Brisse S, Lecuit M. 2017. Real-Time Whole-Genome Sequencing for Surveillance of *Listeria monocytogenes*, France. Emerg Infect Dis 23:1462-1470.

Orsi RH, Borowsky ML, Lauer P, Young SK, Nusbaum C, Galagan JE, Birren BW, Ivy RA, Sun Q, Graves LM, Swaminathan B, Wiedmann M. 2008. Short-term genome evolution of *Listeria monocytogenes* in a non-controlled environment. BMC Genom 9:539.

Orsi RH, den Bakker HC, Wiedmann M. 2011. *Listeria monocytogenes* lineages: genomics, evolution, ecology, and phenotypic characteristics. Int J Med Microbiol 301:79-96.

Painset A, Björkman JT, Kiil K, Guillier L, Mariet JF, Félix B, Amar C, Rotariu O, Roussel S, Perez-Reche F, Brisse S, Moura A, Lecuit M, Forbes K, Strachan N, Grant K, Møller-Nielsen E, Dallman TJ. 2019. LiSEQ - whole-genome sequencing of a cross-sectional survey of *Listeria monocytogenes* in ready-to-eat foods and human clinical cases in Europe. Microb Genom 5: e000257.

Pearce ME, Alikhan NF, Dallman TJ, Zhou Z, Grant K, Maiden MCJ. 2018. Comparative analysis of core genome MLST and SNP typing within a European Salmonella serovar Enteritidis outbreak. Int J Food Microbiol 274:1-11.

Pietzka A, Allerberger F, Murer A, Lennkh A, Stöger A, Cabal Rosel A, Huhulescu S, Maritschnik S, Springer B, Lepuschitz S, Ruppitsch W, Schmid D. 2019. Whole Genome Sequencing Based Surveillance of *L. monocytogenes* for Early Detection and Investigations of Listeriosis Outbreaks. Front Public Health 7:139.

Pightling AW, Petronella N, Pagotto F. 2015. The *Listeria monocytogenes* core -genome sequence typer (LmCGST): a bioinformatics pipeline for molecular characterization with next generation sequence data. BMC Microbiol 15:224.

Pirone-Davies C, Chen, Y, Pightling A, Ryan G, Wang Y, Yao K, Hoffmann M, Allard MW. 2018. Genes significantly associated with lineage II food isolates of *Listeria monocytogenes*. BMC Genomics 19:708.

Ragon M, Wirth T, Hollandt F, Lavenir R, Lecuit M, Monnier Le A, Brisse S. A new perspective on *Listeria Monocytogenes* evolution. 2008. PLoS Pathog 4:e1000146.

Ratan A. 2009. Assembly algorithms for next generation sequence data. Ph.D. dissertation, The Pennsylvania State University.

Ruppitsch W, Pietzka A, Prior K, Bletz S, Fernandez HL, Allerberger F, Harmsen D, Mellmann A. 2015. Defining and evaluating a core genome MLST scheme for whole genome sequence-based typing of *Listeria monocytogenes*. J Clin Microbiol 53:2869-76.

Ryan S, Begley M, Hill C, Gahan CGM. 2010. A Five-Gene Stress Survival Islet (SSI-1) That Contributes to the Growth of *Listeria Monocytogenes* in Suboptimal Conditions. J Appl Microbiol 109:984-95.

Scallan E, Hoekstra RM, Angulo FJ, Tauxe RV, Widdowson MA, Roy SL, Jones JL, Griffin PM. 2011. Foodborne illness acquired in the United States--major pathogens. Emerg Infect Dis 17:7-15.

Swaminathan B, Barrett T, Hunter SB, Tauxe RV, CDC PulseNet Task Force. 2001. PulseNet: The molecular subtyping network for foodborne bacterial disease surveillance, United States. Emerg Infect Dis 7:382-389.

USDA ERS. 2014. Cost estimates of foodborne illnesses. Economic Re- search Service, US Department of Agriculture, Washington, DC. Available at: https://www.ers.usda.gov/data-products/cost-estimates-of-foodborne-illnesses.aspx.

Wang J, Ray AJ, Hammons SR, Oliver HF. 2015. Persistent and transient *Listeria monocytogenes* strains from retail deli environments vary in their ability to adhere and form biofilms and rarely have inlA premature stop codons. Foodborne Pathog Dis 12:151-8.

Ward TJ, Usgaard T, Evans P. 2010. A targeted multilocus genotyping assay for lineage, serogroup, and epidemic clone typing of *Listeria monocytogenes*. Appl Environ Microbiol 76:6680-4.

Tables:

Table 2.1: Comparison of features present in Haplo-ST and other currently available commercial tools for wgMLST of *Listeria monocytogenes*.

	Current wgMLST-based approaches for strain typing		Haplo-ST
	BioNumerics	Ridom SeqSphere+	
Free availability	✘	✘	✓
Database used for allelic nomenclature	BIGSdb-*Lm*	cgMLST.org	BIGSdb-*Lm*
Creates genomes assemblies of isolates before allele calling	Uses both assembly-based and assembly-free approaches	✓	✘ (assembles genes instead of genomes)
Expandable wgMLST schema	✘	✘	✓
Number of loci used for wgMLST	4804, fixed	1701 (cgMLST) + 1158 (accessory genome), fixed	currently 2554, expandable
Output	wgMLST profiles, minimum spanning tree (MST)	wgMLST profiles, MST	wgMLST profiles, allele sequences for wgMLST profiles, paralogous genes
Cluster analysis features included with software	✓	✓	✘
Types of cluster analysis possible	MST	MST, phylogenetic tree (software only provides concatenated allele sequences)	Dendrograms and phylogenetic trees can be constructed with third-party software
Can be used for strain typing of species other than *Listeria monocytogenes*	✓	✓	✓
Automated curation tools (for assigning allele ID's) provided for new alleles	✓	✓	✘

Table 2.2: Global AMOVA results weighted over all variable loci in the two groups of *Listeria monocytogenes* isolates.

Groups of isolates	**Source of variation**	**Variance components**	**Variation (%)**	**Fixation index**
Natural Environment vs. Poultry Processing plants	Among groups	86.61	9.00	F_{ST} = 0.09002
	Within groups	875.51	91.00	
Persistent vs. Transient	Among groups	32.64	3.82	F_{ST} = 0.03821
	Within groups	821.66	96.18	

Table 2.3: Genes showing significant genetic differentiation between groups of *Listeria monocytogenes* isolates collected from the natural environment vs. poultry processing plants and transient vs. persistent strains, and may be enriched for adaptation and persistence of *Listeria monocytogenes* in poultry processing environments.

Gene name	Gene product (obtained from RefSeq)	Biological Function (obtained from KEGG)
lmo0875	PTS beta-glucoside transporter subunit IIB	Carbohydrate metabolism, Membrane transport
lmo1949	hypothetical protein	Ribosome biogenesis
lmo2650	MFS transporter	Carbohydrate metabolism, Membrane transport
lmo1210	hypothetical protein	Electrochemical potential-driven transporters
lmo0687	hypothetical protein	Peptidase
lmo0694	hypothetical protein	Unknown function
lmo0964	hypothetical protein (thioredoxin)	Oxidative stress, Signaling
lmo2078	hypothetical protein	Transfer RNA biogenesis
lmo2383	monovalent cation/H+ antiporter subunit F	Electrochemical potential-driven transporters
lmo2548	50S ribosomal protein L31	Translation
lmo1776	hypothetical protein	Unknown function
lmo1960	ferrichrome ABC transporter ATP-binding protein	Iron complex transporter
lmo1336	5-formyltetrahydrofolate cyclo-ligase	Metabolism of cofactors and vitamins
lmo2689a	hypothetical protein	Uncharacterized
lmo1294	tRNA delta(2)-isopentenylpyrophosphate transferase	Transfer RNA biogenesis, Biosynthesis of secondary metabolites
lmo2640	hypothetical protein	Metabolism of terpenoids and polyketides, Biosynthesis of secondary metabolites
lmo0360	DeoR family transcriptional regulator	Unknown function
lmo1817	hypothetical protein	Metabolism of cofactors and vitamins
lmo2073	ABC transporter ATP-binding protein	Translation factor
lmo1205	cobalamin biosynthesis protein CbiN	Membrane transport
lmo1464	diacylglycerol kinase	Glycan biosynthesis and metabolism

Figures:

Figure 2.1: Haplo-ST, a tool for wgMLST profiling of *Listeria monocytogenes* from WGS reads.

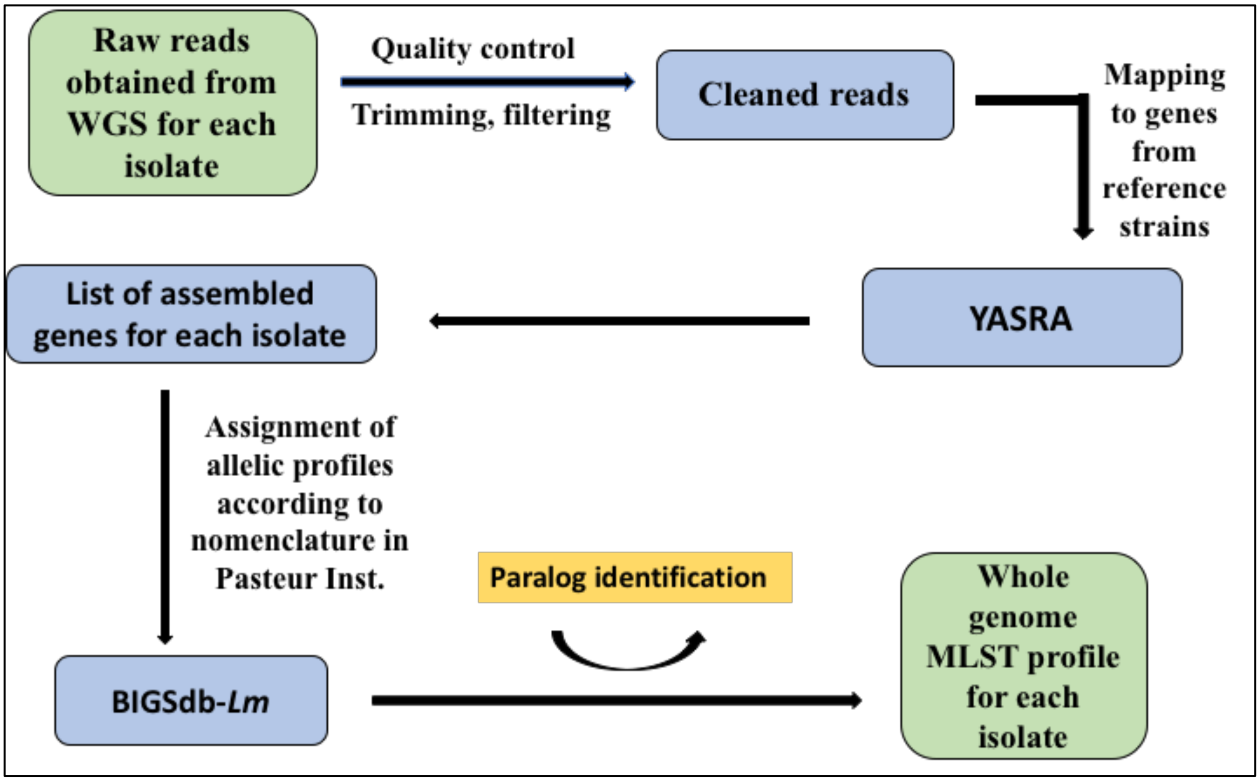
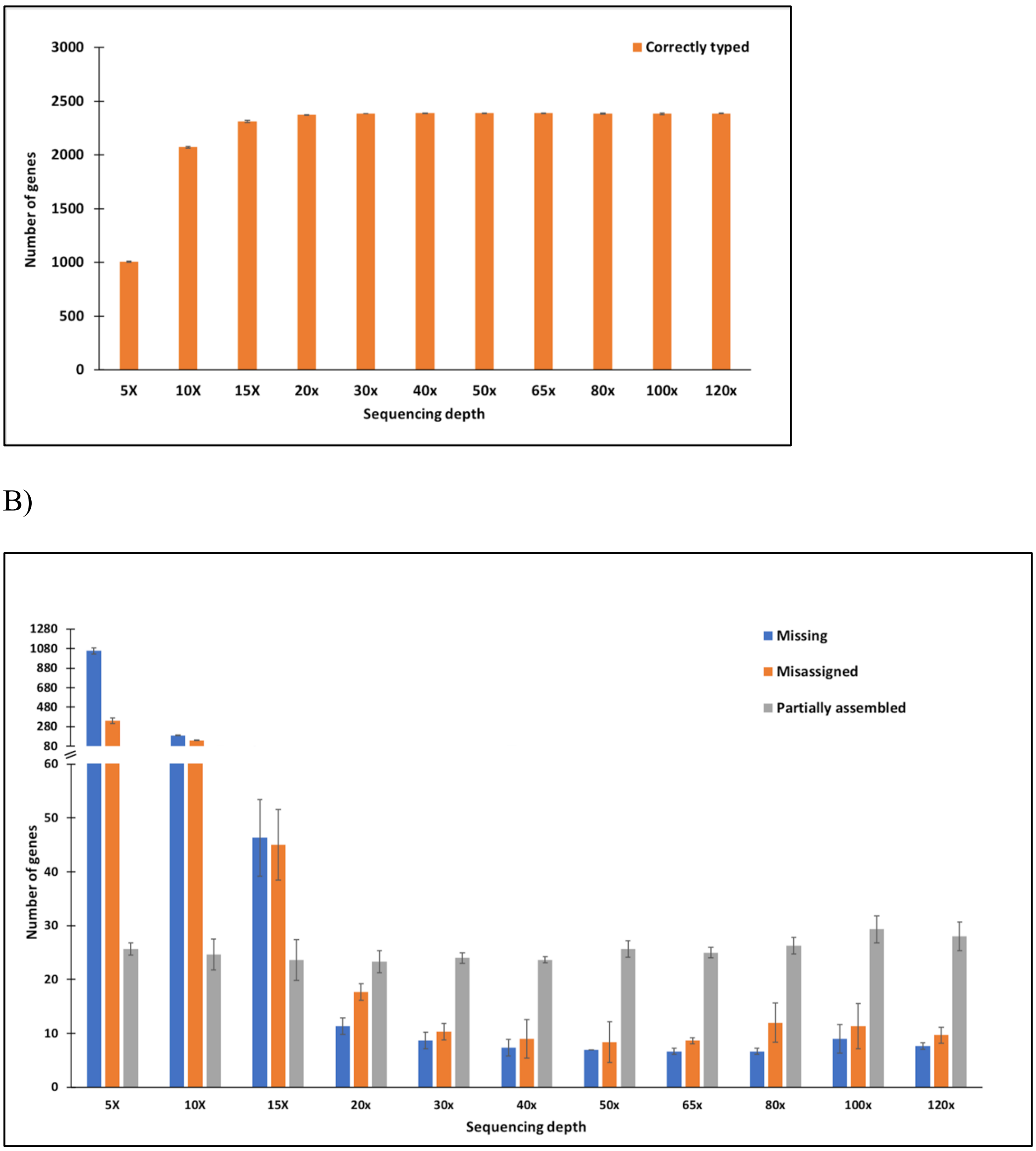

Raw reads obtained from WGS for each isolate
Quality control
Trimming, filtering
Cleaned reads
Mapping to genes from reference strains
YASRA
List of assembled genes for each isolate
Assignment of allelic profiles according to nomenclature in Pasteur Inst.
Paralog identification
BIGSdb-*Lm*
Whole genome MLST profile for each isolate

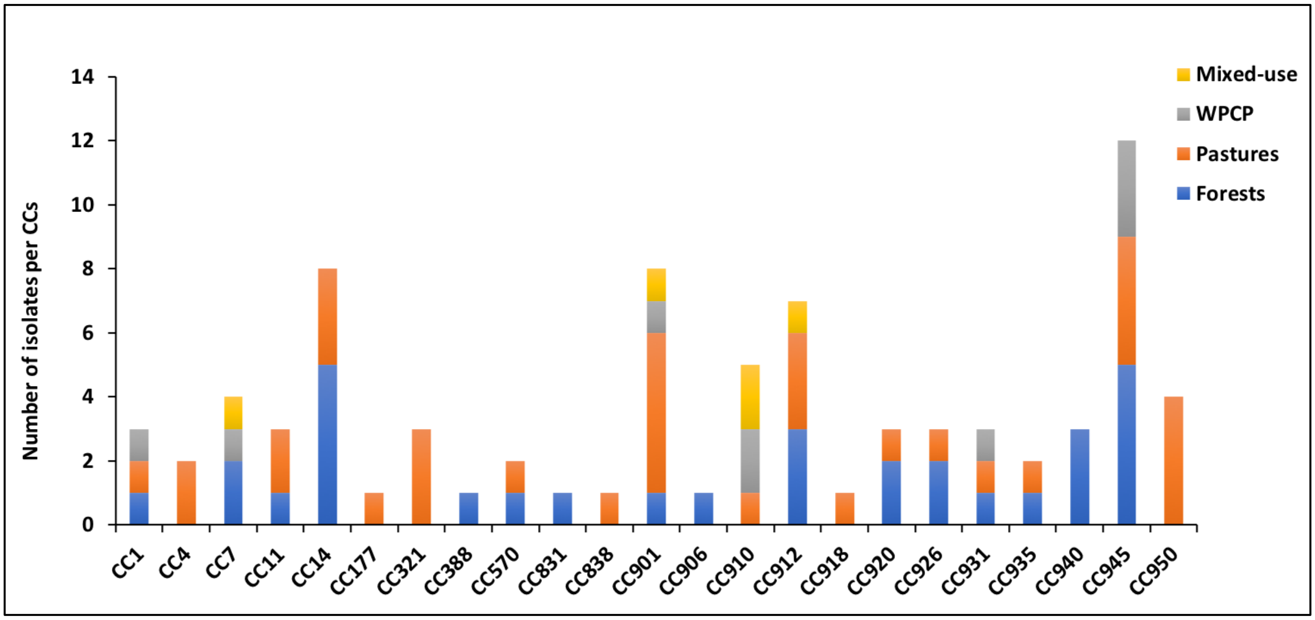
Figure 2.2: Dependency of Haplo-ST on sequencing depth. (A) Simulation of the number of genes correctly profiled by Haplo-ST across sequencing depths ranging from 5× - 120× (B) The number of genes missing an allele ID assignment, the number of genes misassigned an erroneous allele ID and the number of genes partially assembled with Haplo-ST across different sequencing depths.


A)

B)

Figure 2.3: Distribution of CCs in (A) river water flowing through different land use areas (B) *L. monocytogenes* strains isolated from poultry processing plants.

A)



B)

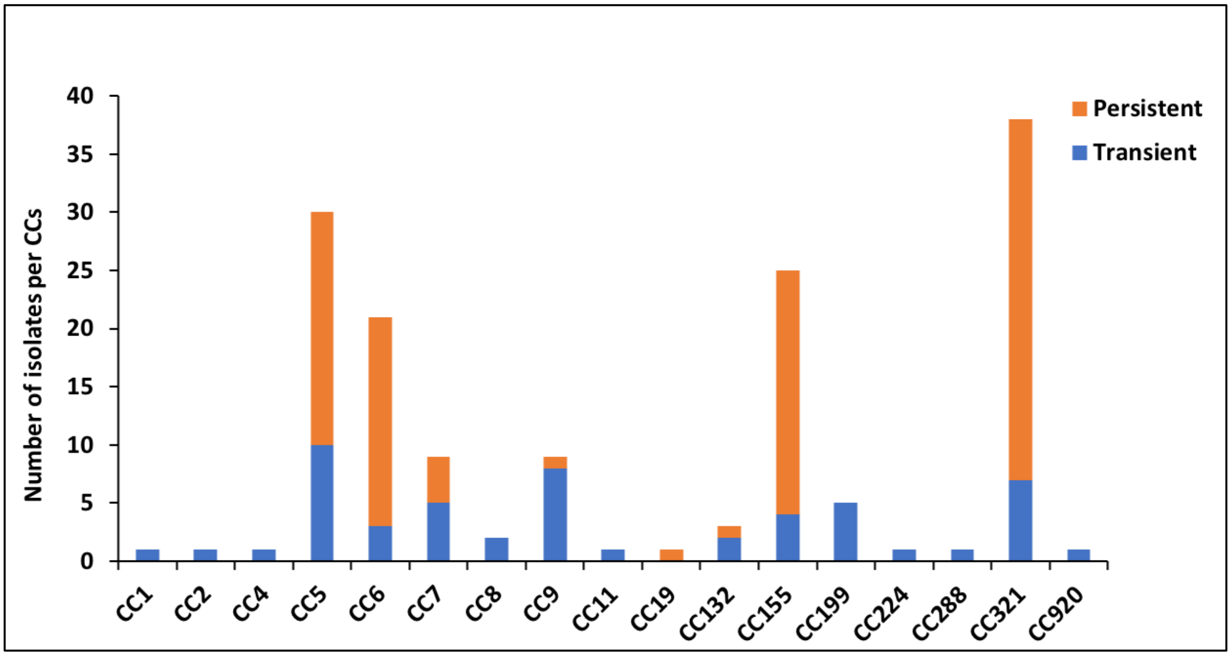


Figure 2.4: Phylogenetic relationships between isolates collected from the natural environment. (A) *Listeria monocytogenes* isolates belonging to lineages I (orange), II (red) and III (blue) form separate clusters in the phylogenetic tree. (B) Random distribution of three lineages of *Listeria monocytogenes* found at different sampling sites.

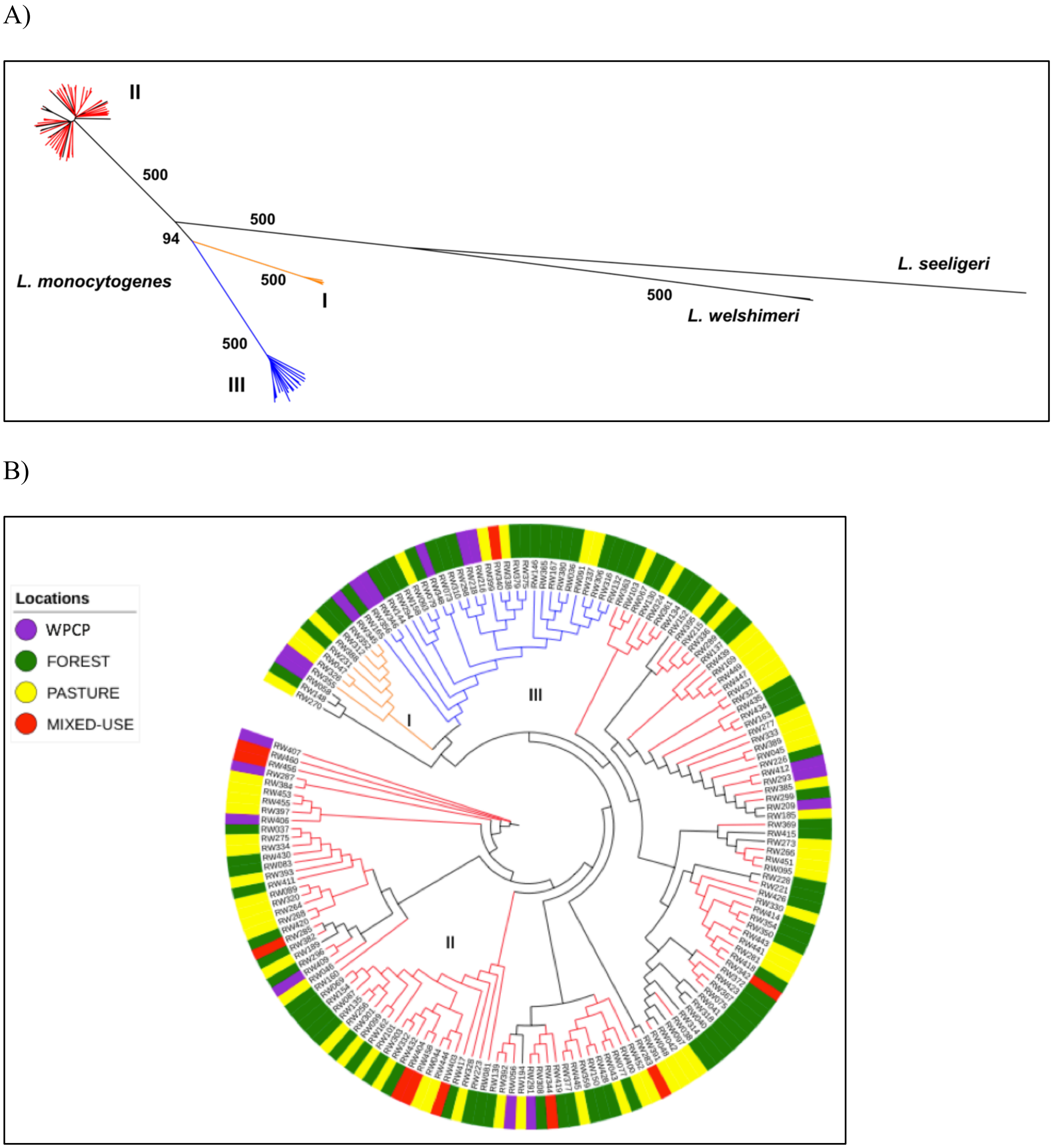
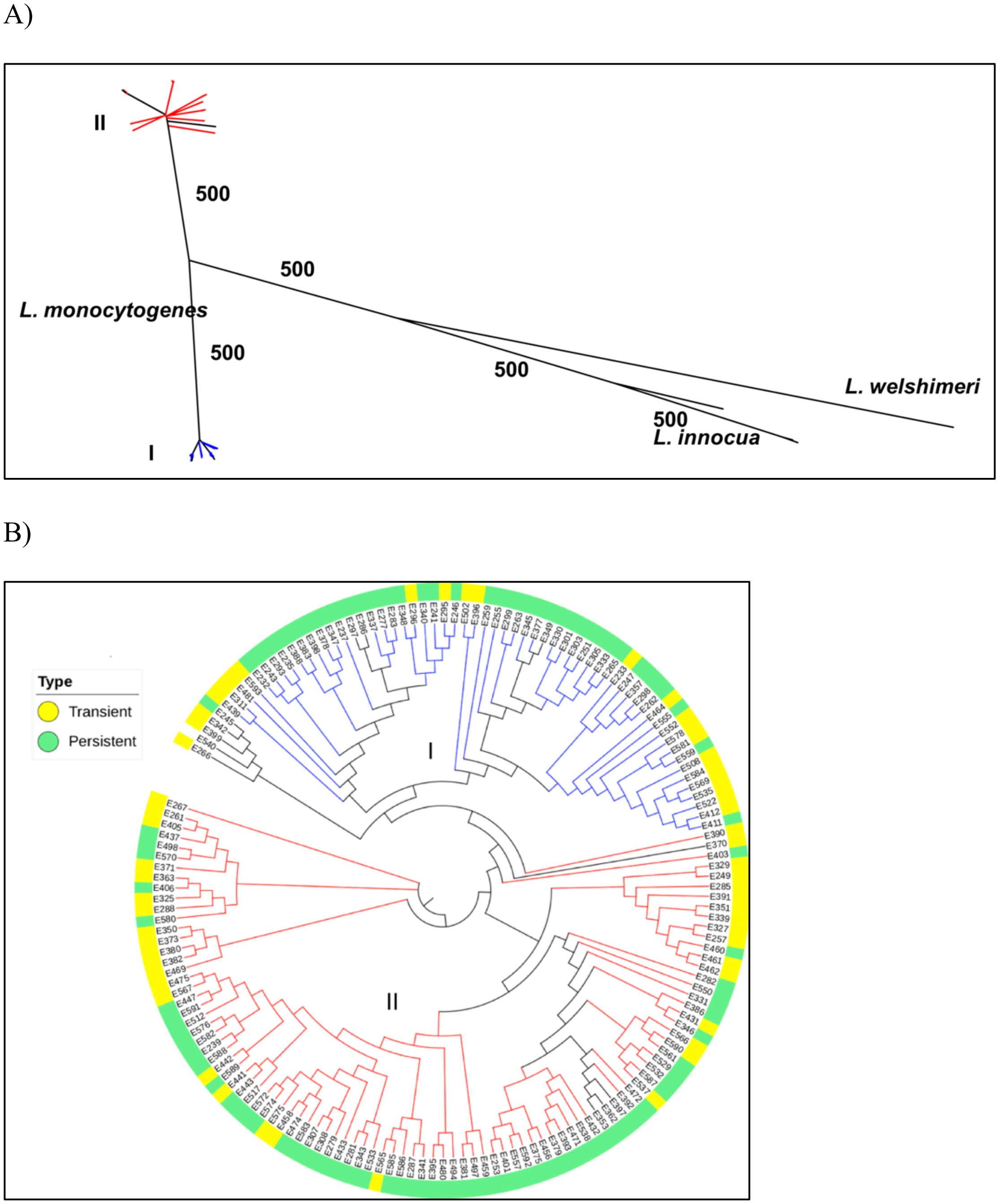
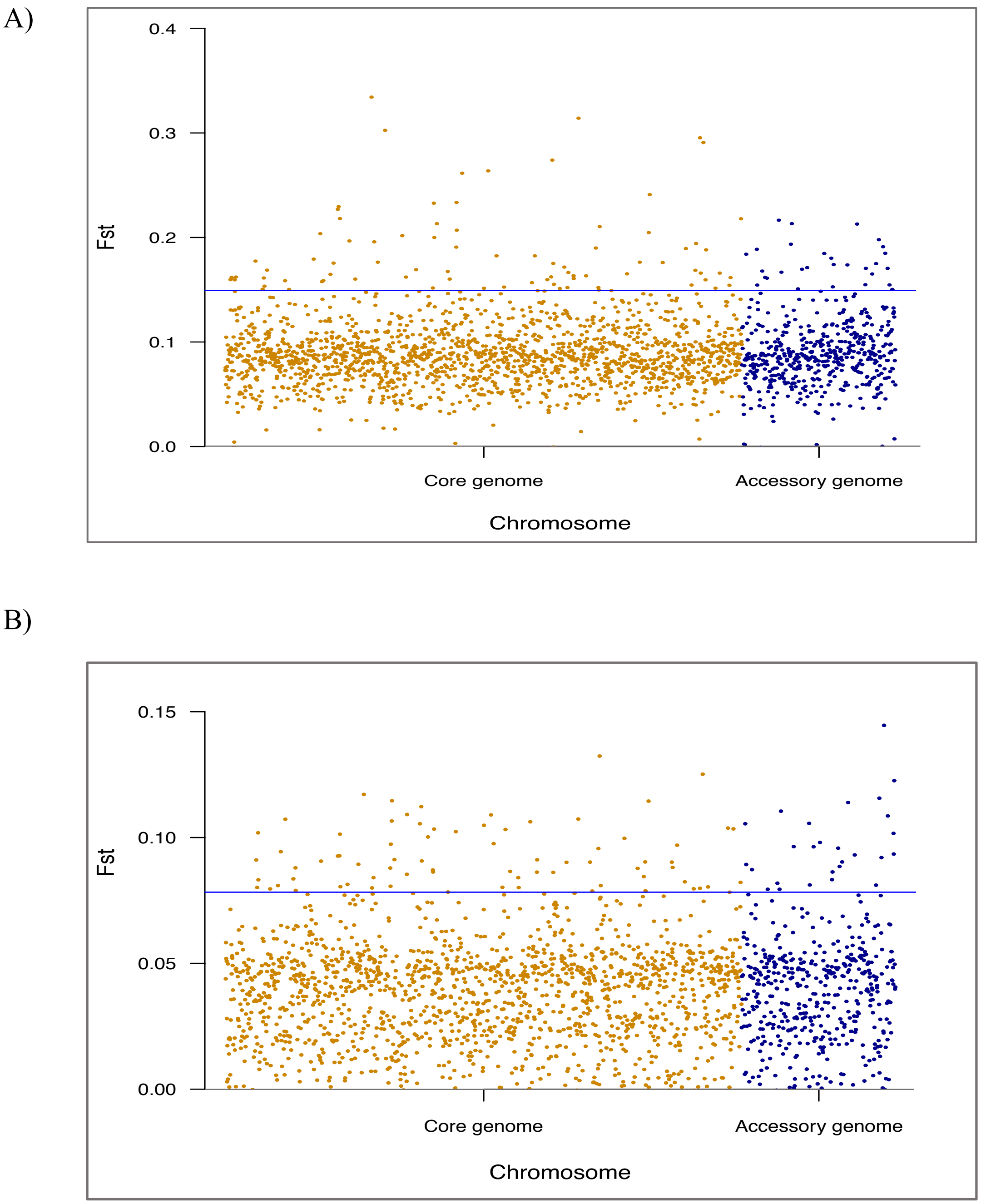


Figure 2.5: Phylogenetic relationships between isolates collected from poultry processing plants. (A) *Listeria monocytogenes* lineages I (blue) and II (red) form two separate groups in the phylogenetic tree, with the majority of isolates belonging to lineage II. (B) Persistent strains were more abundant than transient strains, but there was no lineage-specific association of persistent/transient strains.


A)

B)

Figure 2.6: Manhattan plots of genome-wide F_{ST} values between (A) *Listeria monocytogenes* isolates obtained from the natural environment and poultry processing plants (B) groups of transient and persistent strains. F_{ST} values are shown on the y axis. The loci are arranged in two groups on the x axis; the first group consisting of loci present in the core genome and the other group consisting of loci in the accessory genome as specified in BIGSdb-*Lm.* The significant thresholds (blue line) are set at the top 5% of the F_{ST} distribution.



2.9 SUPPLEMENTAL MATERIAL

Supplemental material is available at:

https://www.dropbox.com/sh/vfau4ysqppe5z3k/AABjYq4i6wO28xJiaCFTANF_a?dl=0

File S1: List of 133 paralogous genes identified in our dataset.

File S2: List of 111 loci with the highest FST values (top 5% of FST distribution) having considerable genetic variation between isolates obtained from the natural environment and poultry processing plants.

File S3: List of 102 loci (upper 5% of FST distribution) having the most genetic divergence between the transient and persistent strains.

File S4: Whole-genome MLST profiles of *L. monocytogenes* isolates generated by Haplo-ST.

File S5: Whole-genome MLST profiles of *L. monocytogenes* isolates with new alleles assigned custom allele-IDs.

File S6: Contains:

> **Figure S1:** Inaccurate characterization by BIGSdb-*Lm* when it recognizes genes that are not paralogous to each other as 'exact matches'.
>
> **Table S1:** Comparison between SNP-based and wgMLST-based approaches for subtyping bacterial strains.

CHAPTER 3

WHOLE GENOME GENETIC VARIATION AND LINKAGE DISEQUILIBRIUM IN A DIVERSE COLLECTION OF *LISTERIA MONOCYTOGENES* ISOLATES[2]

[2] Swarnali Louha, Richard J. Meinersmann, and Travis C. Glenn, Submitted to PLoS One.

3.1 ABSTRACT

We performed whole-genome multi-locus sequence typing for 2554 genes in a large and heterogenous panel of 180 *Listeria monocytogenes* strains having diverse geographical and temporal origins. The subtyping data was used for characterizing genetic variation and evaluating patterns of linkage disequilibrium in the pan-genome of *L. monocytogenes.* Our analysis revealed the presence of strong linkage disequilibrium in *L. monocytogenes*, with ~99% of genes showing significant non-random associations with a large majority of other genes in the genome. Twenty-seven loci having lower levels of association with other genes were considered to be potential "hot spots" for horizontal gene transfer (i.e., recombination via conjugation, transduction, and/or transformation). The patterns of linkage disequilibrium in *L. monocytogenes* suggest limited exchange of foreign genetic material in the genome and can be used as a tool for identifying new recombinant strains. This can help understand processes contributing to the diversification and evolution of this pathogenic bacteria, thereby facilitating development of effective control measures.

3.2 INTRODUCTION

The bacterial genome is a dynamic structure. Characterizing patterns of genomic variation in bacterial pathogens can provide insights into the forces shaping their biology and evolutionary history (Zwick et al 2011). Homologous recombination is an important driver of evolution and increases the adaptive potential of bacteria by allowing variation to be tested across multiple genomic backgrounds (Yahara et al. 2015). Recombination is mediated by three mechanisms; transformation, transduction, and conjugation, and the availability and efficacy of these mechanisms and their biological consequences play a major role in determining the frequency of recombination in a bacterial population (Feil and Spratt 2001, Zwick et al 2011). Recombination is variably distributed in bacterial genomes, with some sites in the genome recombining at a higher or lower frequency than the genomic average, known as hot spots and cold spots respectively (Steiner and Smith 2005). Evidence for recombination and its effect on genomic variation can be obtained by detecting patterns of non-random association of genotypes at different loci within a given population, termed as linkage disequilibrium (Feil and Spratt 2001, Zwick et al 2011). Various methods for detecting linkage disequilibrium have been used to study the extent of genetic recombination shaping the population structures of several bacterial species (Smith et al. 1993, Zwick et al. 2011, Takuno et al. 2012, Vigué and Eyre-Walker 2019).

Listeria monocytogenes, known for causing life-threatening infections in animals and human populations at risk, is one of the bacterial species having the lowest rate of homologous recombination. Genetic diversity in this species is mainly driven by the accumulation of mutations over time, with alleles five times more likely to change by mutation than by recombination (Ragon et al. 2008). *L. monocytogenes* is generally considered to have a clonal genetic structure (Piffaretti et al. 1989, Wiedmannn et al. 1997). The population structure of this bacteria consists of 4 evolutionary lineages (I, II, III and IV) and recombination has been observed between isolates of different lineages; suggesting that although recombination is rare in *L. monocytogenes*, this species is not completely clonal (den Bakker et al. 2008, Dunn et al. 2009, Ragon et al. 2008). Interestingly, homologous recombination is not equally frequent among isolates of different lineages, with lineages II, III and IV showing higher rates of recombination and lower degree of sequence similarity than lineage I (Meinersmann et al. 2004, den Bakker et al. 2008, Orsi et al. 2008, Kuenne et al. 2013).

Whole-genome sequencing studies have shown that *L. monocytogenes* genomes are highly syntenic in their gene content and organization, with a majority of gene-scale differences occurring in the accessory genome and accumulated in a few hypervariable hotspots, prophages, transposons, scattered unique genes and genetic islands encoding proteins of unknown functions (Nelson et al. 2004, Hain et al. 2007, den Bakker et al. 2010, 2013, Kuenne et al. 2013). Several other studies have detected evidence of recombination using a few genes (den Bakker et al. 2008, Cantinelli et al. 2013, Ragon et al. 2008) and indicated the presence of significant linkage disequilibrium in *L. monocytogenes* (Call et al. 2003, Salcedo et al. 2003). However, these studies used a limited number of *L. monocytogenes* isolates and evaluated recombination present in a small fraction of the genome, mostly made up of house-keeping genes, which are assumed to be under negative selection and less subject to homologous recombination.

Prior to the advent of next-generation sequencing technologies, multi locus enzyme electrophoresis (MLEE), was used for generating large data sets for the statistical analysis of bacterial populations. MLEE differentiates organisms by assessing the relative electrophoretic mobilities of intracellular enzymes and indexes allelic variation in multiple chromosomal genes (Mallik 2014). MLEE has been successfully used for studying the extent of linkage disequilibrium in a variety of bacterial species (Piffaretti et al. 1989, Maynard Smith et al. 1993, O'Rourke and Stevens 1993). With the easy and cheap availability of sequencing data in the last decade, MLEE has been replaced with an analogous technique called MLST (multi locus sequence typing) for subtyping bacterial genomes (Salcedo et al. 2003, Moura et al. 2017). We recently provided an approach that can generate whole-genome MLST (wgMLST) based characterization of *L. monocytogenes* isolates from whole-genome sequencing data (Louha et al. 2020). In this study, we use this wgMLST-based approach for characterizing genomic variation and assessing genome-wide patterns of linkage disequilibrium in a large collection of *L. monocytogenes* isolates obtained from diverse ecological niches.

3.3 MATERIALS AND METHODS

3.3.1 *L. monocytogenes* isolate selection

We selected a large and diverse panel of 180 *L. monocytogenes* isolates collected from different ecological communities (File S1). This set included (i) 20 isolates each from food, food contact surfaces (FCS), manure, milk, clinical cases, soil, and ready-to-eat (RTE) products obtained from the NCBI Pathogen Detection database and, (ii) 20 isolates from water and sediment samples in the South Fork Broad River watershed located in Northeast Georgia and 20 isolates from effluents from poultry processing plants (EFPP), provided by the USDA and FSIS.

3.3.2 Whole-genome multi-locus sequence typing (wgMLST)

Whole-genome sequencing data for the 180 *L. monocytogenes* isolates were processed using Haplo-ST (Louha et al. 2020) for allelic profiling of 2554 genes per isolate. Illumina whole-genome sequencing reads obtained as previously described (File S1) were trimmed to remove all bases with a Phred quality score of < 20 from both ends and filtered such that 90% of bases in the clean reads had a quality of at least 20. After trimming and filtering, all remaining reads with lengths of < 50 bp were filtered out. The cleaned reads were assembled into allele sequences with YASRA (Ratan 2009) by mapping to reference genes and provided wgMLST profiles with BIGSdb-*Lm* (available at http://bigsdb.pasteur.fr/listeria).

3.3.3 Analysis of Linkage Disequilibrium

First, the raw wgMLST profiles were filtered to remove paralogous loci and genes were ordered according to their genomic position in the *L. monocytogenes* reference strain EGD-e (NCBI accession number NC_003210.1). Next, new alleles not defined in the BIGSdb-*Lm* database and reported as 'closest matches' to existing alleles in BIGSdb-*Lm* were assigned custom allele ID's with in-house Python scripts. The wgMLST profiles were further filtered to retain loci with < 5% missing data. The remaining loci were used to evaluate linkage disequilibrium (LD) between all pairs of loci with Arlequin v3.5.2 (Excoffier and Lischer 2010). LD tests for the presence of significant statistical association between pairs of loci and is based on an exact test. The test procedure is analogous to Fisher's exact test on a two-by-two contingency table but extended to a contingency table of arbitrary size (Slatkin 1994). For each pair of loci, first a contingency table is constructed. The k_1 x k_2 entries of this table are the observed haplotype frequencies, with k_1 and k_2 being the number of alleles at locus 1 and locus 2, respectively. The LD test consists in obtaining the probability of finding a table with the same marginal totals and which has a probability equal or less than that of the observed contingency table. Instead of enumerating all possible contingency tables, a Markov chain is used to explore the space of all possible tables. To start from a random initial position in the Markov chain, the chain is explored for a pre-defined number of steps (the dememorization phase), such as to allow the Markov chain to forget its initial phase and make it independent from its starting point. The *P*-value of the test is then taken as the proportion of the visited tables having a probability smaller or equal to the observed contingency table. In our analysis, we used 100,000 steps of Markov chain to test the *P*-value of the LD test and 10,000 dememorization steps to reach a random initial position on the Markov chain. The significance level of the LD test was set at a *P*-value of 0.05.

3.3.4 Assessment of genetic diversity

Genetic diversity between *L. monocytogenes* isolates collected from the different ecological niches listed as the isolate sources (File S1) was computed with pairwise F_{ST}'s in Arlequin. F_{ST} measures the proportion of the variance in allele frequencies attributable to variation between populations (Charlesworth and Charlesworth 2010) and has a history of being used as a measure of the level of differentiation between populations in population genetics. Fifty thousand permutations were used to test the significance of the genetic distances at a significance level of 0.05.

The AMOVA procedure in Arlequin was used to compute the pairwise differences in allelic content between isolate wgMLST profiles as a matrix of Euclidean squared distances. This distance matrix was used to compute a minimum spanning tree (MST) between all isolates. The MST was visualized and annotated with iTOL v3 (Letunic and Bork 2016). For better visualization, the MST was converted to circular format and annotations for the source of isolates were displayed in outer external rings.

3.4 RESULTS

We performed whole-genome multi locus sequence typing for 180 *L. monocytogenes* isolates obtained from 9 different source populations. For each isolate, allele sequences were assembled for 2554 genes and provided allele ID's based on the unified nomenclature available in the BIGSdb-*Lm* database (File S2). This dataset was filtered to remove 133 paralogous loci identified by Haplo-ST and all loci with > 5% missing data (alleles not assigned ID's by Haplo-ST), and the remaining 2233 loci (File S3) were ordered according to their position in the *L. monocytogenes* reference genome EGD-e. Figure 1 shows the minimum spanning tree of the 180 isolates inferred from allelic differences in the wgMLST profiles. Two results are apparent. First, we see a long branch (red) containing a majority of isolates obtained from soil and manure clustered together, which suggests the origin of these strains from a common ancestor. Interestingly, 3 clinical strains are also found in this cluster. Secondly, a large number of food-related isolates (~51%, obtained from food, FCS, RTE products and EFPP) clustered together in a single branch of the tree (blue) with short branch-lengths to the tips, suggesting that these strains are closely related to each other. Although this is expected, it is interesting to find a few strains obtained from clinical cases, river water and milk in this cluster. The presence of isolates from unrelated ecological communities could be due to the technique used for constructing the dendrogram, which groups isolates based on pairwise differences in allelic content between isolate wgMLST profiles rather than characterizing differences between all variants in nucleotide sequences.

The genetic differentiation test that computes pairwise F_{ST}'s between isolates collected from different ecological communities (Table 3.1) shows that isolates obtained from soil and manure show considerable genetic differentiation from isolates belonging to other communities, with the exception of isolates obtained from clinical cases. Secondly, isolates from the EFPP-RTE pairing has lower F_{ST} than EFPP pairing from all other locations. Thirdly, the clustering dendrogram (Fig 3.1) and F_{ST} test are supportive of each other in that isolates from RTE, FCS and food are not distinguished as separate populations.

We investigated LD between pairs of genes in the genome using an exact test, which measures non-random associations between alleles at two loci based on the difference between observed and expected allele frequencies. As expected, most gene pairs (~97%) in the genome of *L. monocytogenes* show significant LD among pairs of alleles (Fig 3.2, File S4). A majority of genes (2205 of 2233, ~99%) were found to be at LD with at least 90% of other genes in the genome (File S5). Of the remaining 27 genes (~1%) that were at LD with < 90% of genes (Table 3.2), 10 genes were found to be at LD with < 50% of genes. A single locus, *lmo0046*, was at LD with only 19 other genes.

3.5 DISCUSSION

Our dataset reveals the presence of strong LD in the genome of *L. monocytogenes*. Among the 2233 genes tested for LD, 2205 genes (approx. 99%) were found to have pairwise LD with a majority of other genes (90%) in the genome. High levels of LD can not only arise in highly clonal bacterial populations with low rates of recombination, but may also be temporarily present in bacteria with 'epidemic' population structures, in which high recombination rates randomize association between alleles, but adaptive clones emerge and diversify over the short-term (Smith et al. 1993, Feil and Spratt 2001). Because *Listeria* has a clonal genetic structure, it is difficult to see how this high level of LD can arise except as a consequence of low rates of recombination. This is consistent with studies which report recombination in chromosomal genes as an infrequent event in natural populations of *L. monocytogenes* (Piffaretti et al. 1989, Ragon et al. 2008). Because the extent of genetic linkage is a useful index to the horizontal transfer occurring within a species and can be presented as direct evidence for recombination (Feil and Spratt 2001), the remaining ~1% of genes (Table 3.2) that were at LD with < 90% of genes can be described as "hot spots" for the gain of horizontally acquired information. The extensive linkage disequilibrium that we describe in *L. monocytogenes* is in sharp contrast to other pathogenic bacteria that are naturally competent for transformation and recombine frequently to give rise to either weakly clonal or panmictic population structures (Duncan et al. 1994, Suerbaum et al. 1998, Al Suwayyid et al. 2018).

The *L. monocytogenes* pan-genome is highly conserved but open to limited acquisition of foreign DNA or genetic variability through evolutionary forces such as mutation, duplication or recombination (Kuenne et al. 2013). Evidence for homologous recombination between closely related strains of *L. monocytogenes* has been detected by multiple studies, however, non-homologous recombination seems to be rare (Orsi et al. 2008, Dunn et al. 2009, Nightingale et al. 2005). Although recombination via conjugation and generalized transduction has been reported in *L. monocytogenes* (Flamm et al. 1984, Lebrun et al. 1992, Hodgson et al. 2000), and most competence related genes are present in all *Listeria* genomes (Buchrieser 2007), natural competence or induced competence under laboratory conditions has not been observed in *L. monocytogenes* (Borezee et al. 2000, Glaser et al. 2001). This lack of competence may partially explain the low levels of gene acquisition from external gene pools. Limited gene acquisition may also be facilitated by defense systems for foreign DNA/mobile elements such as restriction-modification and/or CRISPR systems, both of which have been shown to restrict horizontal gene transfer in other bacterial genera (den Bakker et al. 2010).

The frequency of recombination in *L. monocytogenes* differs considerably in different regions of the genome and between isolates of different lineages (den Bakker et al. 2008, 2013). This may arise from differences in selective pressures in the environment and varying degrees of horizontal gene transfer. Several comparative genomic studies report a clustered distribution of accessory genes on the right replichore of the *L. monocytogenes* genome (approx. 500 Kb in the first 65°), indicating an area of high genome plasticity (Kuenne et al. 2013, den Bakker et al. 2013). On the contrary, a study by Orsi et al. failed to find any evidence of spatial clustering in a large number of genes which show evidence for recombination in *L. monocytogenes* (Orsi et al. 2008). Further, a recent study described the presence of homologous recombination in nearly 60% of loci in the core genome of *L. monocytogenes*, although most of this variation was also found to be affected by purifying selection and was thus neutral (Moura et al. 2016). This is consistent with results from our analysis which finds linkage equilibrium between only ~1% of gene pairs in the genome. Also, genes considered as potential recombination hot spots (Table 3.2) in our dataset are found to be scattered in the genome. A large number (~41%) of these "hot spot" genes (*lmo0046*, *lmo2624*, *lmo2856*, *lmo1469*, *lmo2616*, *lmo1816*, *lmo0248*, *lmo1335*, *lmo2047*, *lmo2628*, *lmo2614*), encode ribosomal proteins and their related subunits. According to the complexity theory (Jain et al. 1999), informational genes involved in complex biosystems and maintenance of basal cellular functions are usually conserved, as they might be less likely to be compatible in the systems of other species. Thus, housekeeping genes such as ribosomal proteins are generally considered to be relatively restricted to horizontal gene transfer. However, several reports suggest horizontal gene transfer of ribosomal proteins in many prokaryotic genomes (Brochier et al. 2000, Makarova et al. 2001, Garcia-Vallve et al. 2002, Chen et al. 2009). Two other "hot spot" genes (*lmo0865*, *lmo2014*) are involved in carbohydrate and amino acid metabolism and have shown evidence for recombination in a prior study (Orsi et al. 2008), indicating that the rapid diversification of these genes may enable *L. monocytogenes* to adapt to environments with varying nutrient availabilities. Some of the other genes encode a variety of internalin's (*lmo0263*, *lmo0514*, *lmo0264*, lmo0434), transporters (*lmo0756*, *lmo1839*), transcriptional regulators (*lmo0659*), cell surface proteins (*lmo2179*), other invasion-associated proteins (*lmo0582*), and proteins involved in response to temperature fluctuations (*lmo1364*, *lmo2206*). Internalin's are cell surface proteins with known and hypothesized roles in virulence (den Bakker et al. 2010, Tsai et al. 2011). Evidence of recombination in internalin's and these other genes suggests that *L. monocytogenes* is subjected to sustained selection pleasures in the environment, and it responds to these pressures by continuously regulating its transcriptional machinery and remodeling the cell surface, thereby facilitating adaptation within the host and as a saprophyte.

In conclusion, we have identified the presence of strong linkage disequilibrium in the genome of *L. monocytogenes.* Parts of the genome showing strong non-random association between genes are highly conserved regions, and are most possibly affected by positive selection. The low levels of recombination within the *L. monocytogenes* genome suggests that the patterns of association observed between genes can be used to recognize newly emerging strains and help in understanding the processes involved in the diversification and evolution of *L. monocytogenes.* Such investigations can ultimately help to develop better control measures for this pathogenic microbe.

3.6 ACKNOWLEDGEMENTS

This research was supported by funding from USDA Agricultural Research Service Project no. 6040-32000-009-00-D. We thank USDA and FSIS for providing us with *Listeria monocytogenes* whole-genome sequencing samples from river water and effluents of poultry processing plants. The high-performance computing cluster at Georgia Advanced Computing Resource Center (GACRC) at the University of Georgia provided computational infrastructure and technical support throughout the work.

3.7 REFERENCES

Al Suwayyid BA, Coombs GW, Speers DJ, Pearson J, Wise MJ, Kahler CM. 2018. Genomic epidemiology and population structure of *Neisseria gonorrhoeae* from remote highly endemic Western Australian populations. BMC Genomics 19:165.

Borezee E, Msadek T, Durant L, Berche P. 2000. Identification in *Listeria monocytogenes* of MecA, a homologue of the *Bacillus subtilis* competence regulatory protein. J Bacteriol 182:5931-5934.

Brochier C, Philippe H, Moreira D. 2000. The evolutionary history of ribosomal protein RpS14: horizontal gene transfer at the heart of the ribosome. Trends Genet 16:529-533.

Buchrieser C. 2007. Biodiversity of the species *Listeria monocytogenes* and the genus Listeria. Microbes Infect 9:1147-1155.

Call DR, Borucki MK, Besser TE. 2003. Mixed-genome Microarrays Reveal Multiple Serotype and Lineage-Specific Differences Among Strains of *Listeria monocytogenes.* J Clin Microbiol 41:632-639.

Cantinelli T, Chenal-Francisque V, Diancourt L, Frezal L, Leclercq A, Wirth T, Lecuit M, Brisse S. 2013. "Epidemic clones" of *Listeria monocytogenes* are widespread and ancient clonal groups. J Clin Microbiol 51:3770-3779.

Charlesworth B, Charlesworth D. 2010. Elements of Evolutionary Genetics, Roberts and Company publishers, Greenwood Village, Colorado.

Chen K, Roberts E, Luthey-Schulten Z. 2009. Horizontal gene transfer of zinc and non-zinc forms of bacterial ribosomal protein S4. BMC Evol Biol 9:179.

den Bakker HC, Cummings CA, Ferreira V, Vatta P, Orsi RH, Degoricija L, Barker M, Petrauskene O, Furtado MR, Wiedmann M. 2010. Comparative genomics of the bacterial genus Listeria: Genome evolution is characterized by limited gene acquisition and limited gene loss. BMC Genomics 11:688.

den Bakker HC, Desjardins CA, Griggs AD, Peters JE, Zeng Q, Young SK, Kodira CD, Yandava C, Hepburn TA, Haas BJ, Birren BW, Wiedmann M. 2013. Evolutionary Dynamics of the Accessory Genome of *Listeria monocytogenes*. PLoS One 8:e67511.

den Bakker HC, Didelot X, Fortes ED, Nightingale KK, Wiedmann M. 2008. Lineage Specific Recombination Rates and Microevolution in *Listeria monocytogenes*. BMC Evol Biol 8:277.

Duncan KE, Ferguson N, Kimura K, Zhou X, Istock CA. 1994. Fine-scale genetic and phenotypic structure in natural populations of *Bacillus subtilis* and *Bacillus licheniformis*: implications for bacterial evolution and speciation. Evolution 48:2002-2025.

Dunn KA, Bielawski JP, Ward TJ, Urquhart C, Gu H. 2009. Reconciling Ecological and Genomic Divergence Among Lineages of *Listeria* Under an "Extended Mosaic Genome Concept". Mol Biol Evol 26:2605-2615.

Excoffier L, Lischer HEL. 2010. Arlequin suite ver 3.5: A new series of programs to perform population genetics analyses under Linux and Windows. Mol Ecol Resour 10:564-567.

Feil EJ, Spratt BG. 2001. Recombination and the population structures of bacterial pathogens. Annu Rev Microbiol 55:561-590.

Flamm RK, Hinrichs DJ, Thomashow MF. 1984. Introduction of pAM beta 1 into *Listeria monocytogenes* by conjugation and homology between native *L. monocytogenes* plasmids. Infect Immun 44:157-161.

Garcia-Vallve S, Simo FX, Montero MA, Arola L, Romeu A. 2002. Simultaneous horizontal gene transfer of a gene coding for ribosomal protein l27 and operational genes in *Arthrobacter* sp. J Mol Evol 55:632-637.

Glaser P, Frangeul L, Buchrieser C, Rusniok C, Amend A, Baquero F, Berche P, Bloecker H, Brandt P, Chakraborty T, Charbit A, Chetouani F, Couvé E, de Daruvar A, Dehoux P, Domann E, Domínguez-Bernal G, Duchaud E, Durant L, Dussurget O, Entian KD, Fsihi H, García-del Portillo F, Garrido P, Gautier L, Goebel W, Gómez-López N, Hain T, Hauf J, Jackson D, Jones LM, Kaerst U, Kreft J, Kuhn M, Kunst F, Kurapkat G, Madueno E, Maitournam A, Vicente JM, Ng E, Nedjari H, Nordsiek G, Novella S, de Pablos B, Pérez-Diaz JC, Purcell R, Remmel B, Rose M, Schlueter T, Simoes N, Tierrez A, Vázquez-Boland JA, Voss H, Wehland J, Cossart P. 2001. Comparative genomics of *Listeria* species. Science 294:849-852.

Hain T, Chatterjee SS, Ghai R, Kuenne CT, Billion A, Steinweg C, Domann E, Kärst U, Jänsch L, Wehland J, Eisenreich W, Bacher A, Joseph B, Schär J, Kreft J, Klumpp J, Loessner MJ, Dorscht J, Neuhaus K, Fuchs TM, Scherer S, Doumith M, Jacquet C, Martin P, Cossart P, Rusniock C, Glaser P, Buchrieser C, Goebel W, Chakraborty T. 2007. Pathogenomics of *Listeria* spp. Int J Med Microbiol 297:541-557.

Hodgson DA. 2000. Generalized transduction of serotype 1/2 and serotype 4b strains of *Listeria monocytogenes*. Mol. Microbiol 35:312-323.

Jain R, Rivera MC, Lake JA. 1999. Horizontal gene transfer among genomes: the complexity hypothesis. Proc Natl Acad Sci USA 96:3801-3806.

Kuenne C, Billion A, Mraheil MA, Strittmatter A, Daniel R, Goesmann A, Barbuddhe S, Hain T, Chakraborty T. 2013. Reassessment of the *Listeria monocytogenes* Pan-Genome Reveals Dynamic Integration Hotspots and Mobile Genetic Elements as Major Components of the Accessory Genome. BMC Genomics 14:47.

Lebrun M, Loulergue J, Chaslus-Dancla E, Audurier A. 1992. Plasmids in *Listeria monocytogenes* in relation to cadmium resistance. Appl Environ Microbiol 58:3183-3186.

Letunic I, Bork P. 2016. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic Acids Res 44:W242-245.

Louha S, Meinersmann RJ, Abdo Z, Berrang ME, Glenn TC. 2020. An Open-Source Program (Haplo-ST) for Whole-Genome Sequence Typing shows Extensive Diversity of *Listeria monocytogenes* in Outdoor Environments and Poultry Processing Plants. Appl Environ Microbiol 87(1). DOI: 10.1128/AEM.02248-20

Makarova KS, Ponomarev VA, Koonin EV. 2001. Two C or not two C: recurrent disruption of Zn-ribbons, gene duplication, lineage-specific gene loss, and horizontal gene transfer in evolution of bacterial ribosomal proteins. Genome Biol 2:RESEARCH 0033.

Mallik S. 2014. IDENTIFICATION METHODS | Multilocus Enzyme Electrophoresis, p 336-343. *In* Batt CA, Tortorello ML (ed), Reference Module in Food Science: Encyclopedia of Food Microbiology, Second Edition.

Meinersmann RJ, Phillips RW, Wiedmann M, Berrang ME. 2004. Multilocus sequence typing of *Listeria monocytogenes* by use of hypervariable genes reveals clonal and recombination histories of three lineages. Appl Environ Microbiol 70:2193-2203.

Moura A, Criscuolo A, Pouseele H, Maury MM, Leclercq A, Tarr C, Björkman JT, Dallman T, Reimer A, Enouf V, Larsonneur E, Carleton H, Bracq-Dieye H, Katz LS, Jones L, Touchon M, Tourdjman M, Walker M, Stroika S, Cantinelli T, Chenal-Francisque V, Kucerova Z, Rocha EP, Nadon C, Grant K, Nielsen EM, Pot B, Gerner-Smidt P, Lecuit M, Brisse S. 2016. Whole genome-based population biology and epidemiological surveillance of *Listeria monocytogenes*. Nat Microbiol 2:16185.

Nelson KE, Fouts DE, Mongodin EF, Ravel J, DeBoy RT, Kolonay JF, Rasko DA, Angiuoli SV, Gill SR, Paulsen IT, Peterson J, White O, Nelson WC, Nierman W, Beanan MJ, Brinkac LM, Daugherty SC, Dodson RJ, Durkin AS, Madupu R, Haft DH, Selengut J, Van Aken S, Khouri H, Fedorova N, Forberger H, Tran B, Kathariou S, Wonderling LD, Uhlich GA, Bayles DO, Luchansky JB, Fraser CM. 2004. Whole Genome Comparisons of Serotype 4b and 1/2a Strains of the Food-Borne Pathogen *Listeria monocytogenes* Reveal New Insights Into the Core Genome Components of This Species. Nucleic Acids Res 32:2386-2395.

Nightingale K, Windham K, Wiedmann M. 2005. Evolution and molecular phylogeny of *Listeria monocytogenes* isolated from human and animal listeriosis cases and foods. J Bacteriol 187:5537-5551.

O'Rourke M, Stevens E. 1993. Genetic structure of *Neisseria gonorrhoeae* populations : a non-clonal pathogen. J Gen Microbiol 139:2603-2611.

Orsi RH, Sun Q, Wiedmann M. 2008. Genome-wide analyses reveal lineage specific contributions of positive selection and recombination to the evolution of *Listeria monocytogenes*. BMC Evol Biol 8:233.

Piffaretti JC, Kressebuch H, Aeschbacher M, Bille J, Bannerman E, Musser JM, Selander RK, Rocourt J. 1989. Genetic characterization of clones of the bacterium *Listeria monocytogenes* causing epidemic disease. Proc Natl Acad Sci USA 86:3818-3822.

Ragon M, Wirth T, Hollandt F, Lavenir R, Lecuit M, Le Monnier A, Brisse S. 2008. A new perspective on *Listeria monocytogenes* evolution. PLoS Pathog 4:e1000146.

Ratan A. 2009. Assembly algorithms for next generation sequence data. Ph.D. Dissertation, The Pennsylvania State University. Available from: https://etda.libraries.psu.edu/files/final_submissions/587

Salcedo C, Arreaza L, Alcalá B, de la Fuente L, Vázquez JA. 2003. Development of a multilocus sequence typing method for the analysis of *Listeria monocytogenes* clones. J Clin Microbiol 41:757-762.

Slatkin M. 1994. Linkage disequilibrium in growing and stable populations. Genetics 137:331-336.

Smith JM, Smith NH, O'Rourke M, Spratt BG. 1993. How clonal are bacteria? Proc Natl Acad Sci USA 90:4384-4388.

Steiner WW, Smith GR. 2005. Natural meiotic recombination hot spots in the *Schizosaccharomyces pombe* genome successfully predicted from the simple sequence motif M26. Mol Cell Biol 25:9054-9062.

Suerbaum S, Smith JM, Bapumia K, Morelli G, Smith NH, Kunstmann, Dyrek I, Achtman M. 1998. Free recombination within *Helicobacter pylori*. Proc Natl Acad Sci USA 95:12619-12624.

Takuno S, Kado T, Sugino RP, Nakhleh L, Innan H. 2012. Population Genomics in Bacteria: A Case Study of *Staphylococcus aureus*. Mol Biol Evol 29:797-809.

Vigué L, Eyre-Walker A. 2019. The comparative population genetics of *Neisseria meningitidis* and *Neisseria gonorrhoeae*. PeerJ 7:e7216.

Wiedmann M, Bruce JL, Keating C, Johnson AE, McDonough PL, Batt CA. 1997. Ribotypes and Virulence Gene Polymorphisms Suggest Three Distinct *Listeria monocytogenes* Lineages With Differences in Pathogenic Potential. Infect Immun 65:2707-2716.

Yahara K, Didelot X, Jolley KA, Kobayashi I, Maiden MC, Sheppard SK, Falush D. 2016. The Landscape of Realized Homologous Recombination in Pathogenic Bacteria. Mol Biol Evol 33:456-471.

Zwick ME, Thomason MK, Chen PE, Johnson HR, Sozhamannan S, Mateczun A, Read TD. 2011. Genetic variation and linkage disequilibrium in *Bacillus anthracis*. Sci Rep 1:169.

Tables:

Table 3.1: Pairwise genetic distances (FST) between groups of *L. monocytogenes* strains isolated from nine different ecological niches.

	clinical	**food**	**FCS**	**manure**	**milk**	**RTE product**	**soil**	**River water**
clinical	0							
food	0.051*	0						
FCS	0.062*	0.015	0					
manure	0.067*	0.126*	0.137*	0				
milk	0.047*	0.047*	0.073*	0.124*	0			
RTE product	0.09*	0.004	0.007	0.159*	0.069*	0		
soil	0.064*	0.11*	0.124*	0.019*	0.104*	0.135*	0	
River water	0.094*	0.091*	0.107*	0.153*	0.069*	0.092*	0.113*	0
EFPP	0.165*	0.157*	0.137*	0.221*	0.146*	0.076*	0.189*	0.13*

(*P < 0.05)

Table 3.2: Genes at LD with < 90% of genes in the genome of *L. monocytogenes*, thus showing significant evidence for horizontal genetic transfer.

Gene name	Number of genes at LD	Percentage of genes at LD	Location in the chromosome of EGD-e (bp)	Location in core/accessory genome*	Function
lmo0046	19	0.85	50514..50753	core	small subunit ribosomal protein S18
lmo2624	185	8.289	2701254..2701445	core	large subunit ribosomal protein L29
lmo2856	215	9.63	2943569..2943703	accessory	large subunit ribosomal protein L34
lmo1364	239	10.71	1387014..1387214	accessory	Cold shock protein
lmo1469	454	20.34	1501881..1502054	core	small subunit ribosomal protein S21
lmo2616	458	20.52	2697988..2698347	accessory	large subunit ribosomal protein L18
lmo1816	484	21.69	1890951..1891139	core	large subunit ribosomal protein L28
lmo0248	576	25.81	265029..265454	accessory	large subunit ribosomal protein L11
lmo1335	880	39.43	1363826..1363975	core	large subunit ribosomal protein L33
inlH (lmo0263)	1006	45.07	284365..286011	accessory	internalin H
cwhA (lmo0582)	1223	54.79	618932..620380	accessory	Invasion associated secreted endopeptidase
lmo2047	1377	61.69	2130228..2130401	accessory	large subunit ribosomal protein L32

lmo2628	1508	67.56	2702909..2703187	accessory	small subunit ribosomal protein S19
lmo2614	1580	70.79	2697267..2697446	core	large subunit ribosomal protein L30
lmo0758	1606	71.95	783901..784788	core	Hypothetical protein
lmo0514	1699	76.12	547520..549337	accessory	Internalin
lmo0305	1709	76.57	329923..330999	core	L-allo-threonine aldolase
lmo0659	1771	79.35	699410..700306	accessory	Transcriptional regulator
lmo2206	1791	80.24	2294555..2297155	accessory	Heat shock proteins
lmo0756	1797	80.51	781896..782801	core	ABC Transporters
lmo0865	1859	83.29	903837..905510	core	Amino sugar and nucleotide sugar metabolism
lmo2014	1888	84.59	2088797..2091454	accessory	Glycan biosynthesis and metabolism
lmo1611	1904	85.3	1654902..1655975	core	Aminopeptidase
inlE (lmo0264)	1913	85.71	286219..287718	accessory	Internalin E
lmo1839	1925	86.25	1916166..1917452	accessory	Electrochemical potential-driven transporters
lmo2179	1968	88.17	2264772..2268230	accessory	Peptidoglycan binding protein
inlB (lmo0434)	1981	88.75	457021..458913	accessory	Internalin B

Figures:

Figure 3.1: Patterns of genetic differentiation in the 180 *L. monocytogenes* isolates. Minimum spanning tree based on a distance matrix measuring pairwise differences in allelic content between isolate wgMLST profiles. The isolation source of each isolate is indicated with colors on the outer ring. Majority of the isolates sampled from soil and manure cluster together in a distant branch (red), suggesting their recent emergence from a common ancestor. A large number of food-related isolates cluster together in a single branch of the tree (blue), suggesting their close relatedness.

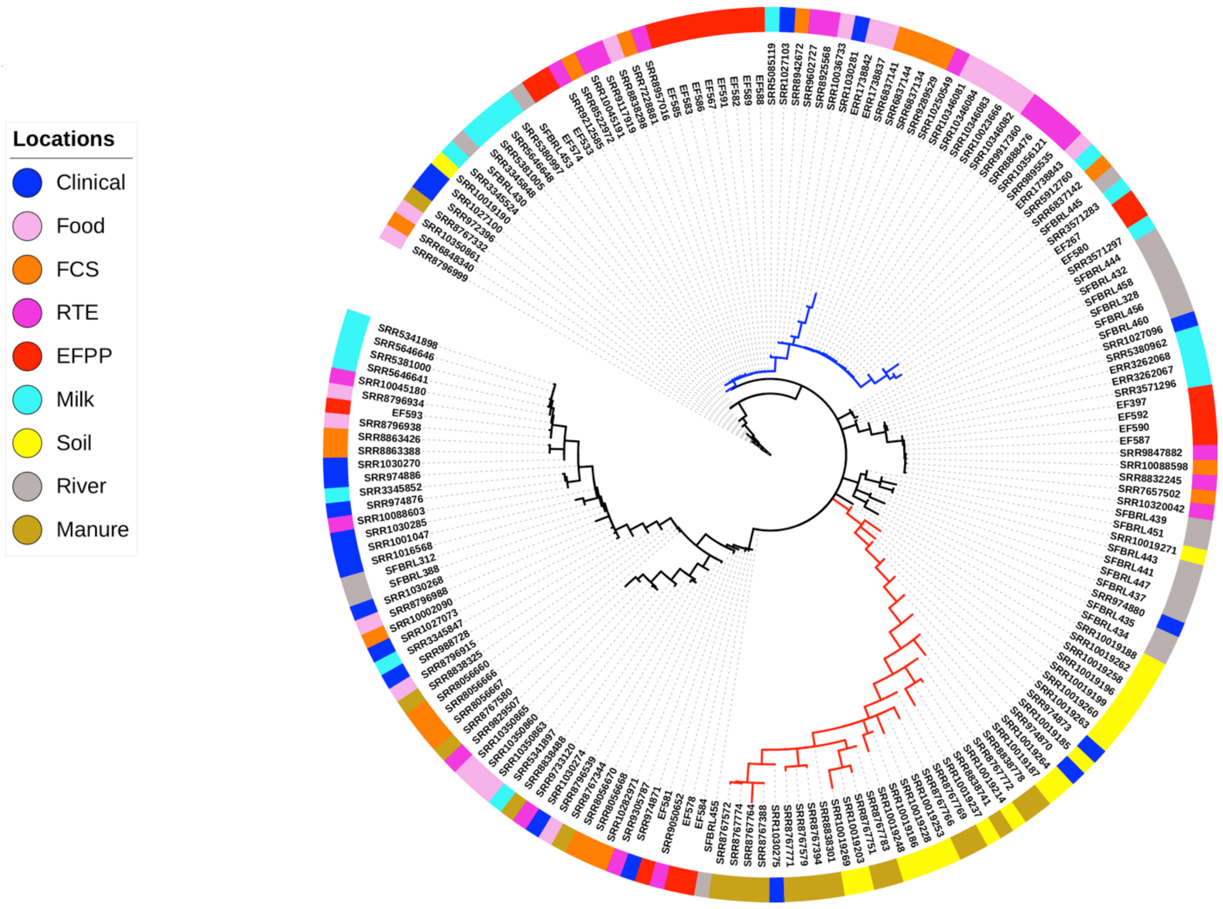
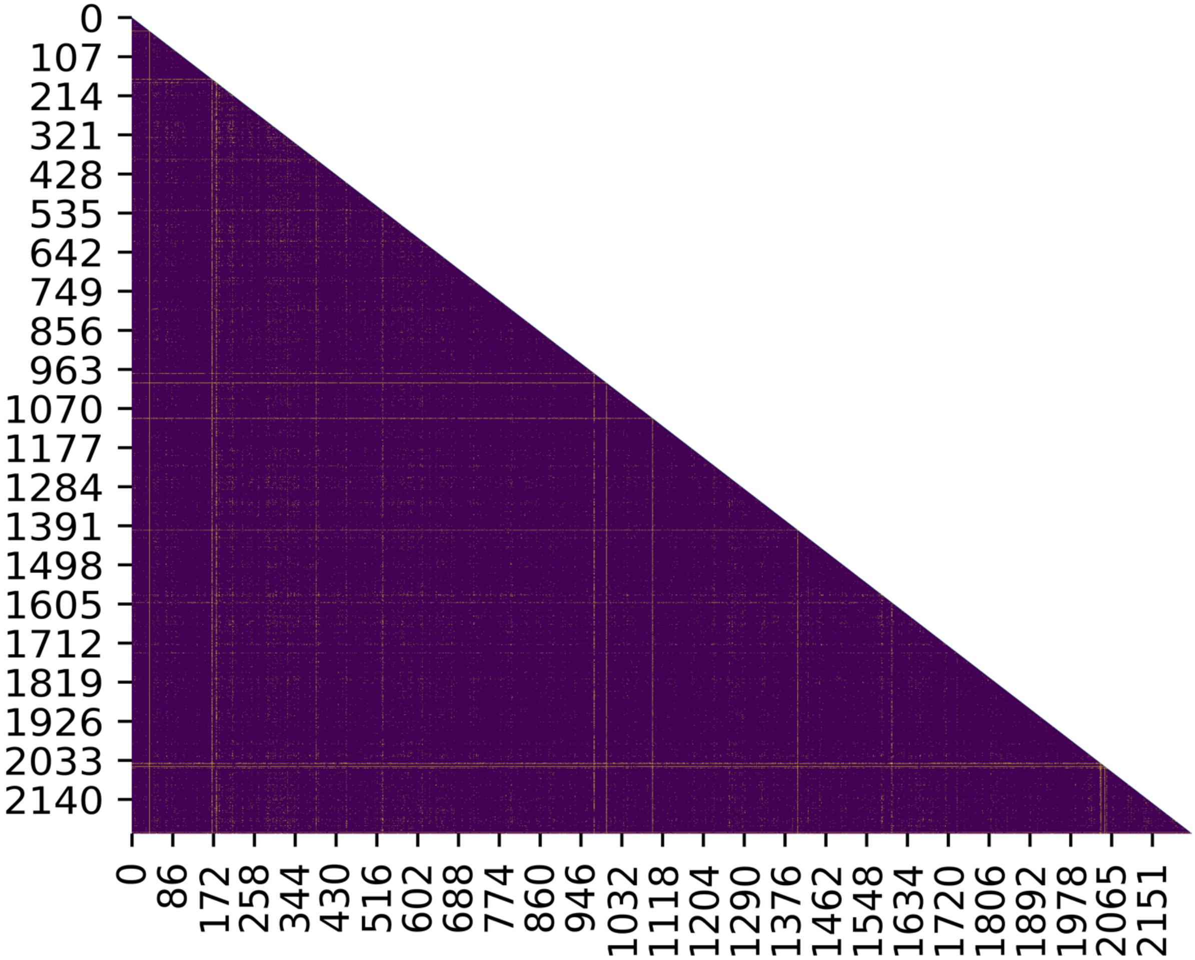


Figure 3.2: Heatmap of the extent of LD in the genome of *L. monocytogenes*. Genes are ordered according to their genomic positions in the *L. monocytogenes* reference strain EGD-e along the *x* and *y* axis (for gene names see File S4). A majority of genes show significant LD in the genome (indigo), while few genes are at linkage equilibrium (yellow).



3.8 SUPPLEMENTAL MATERIAL

Supplemental material is available at:

https://www.dropbox.com/sh/nupygyepmx3v586/AADkG4_yVZj8XoXKHiTkRsTma?dl=0

File S1. Panel of 180 *L. monocytogenes* isolates collected from different ecological communities**.**

File S2. Whole-genome MLST profiles of the 180 *L. monocytogenes* isolates.

File S3. Whole-genome MLST profiles of 2233 loci retained for AMOVA after filtering out paralogous loci and loci with > 5% of missing data.

File S4. Heatmap of LD in the genome of *L. monocytogenes*.

File S5. Percentage of genes at LD with each gene in the genome of *L. monocytogenes*.

CHAPTER 4

A HIGH-QUALITY GENOME ASSEMBLY OF THE NORTH AMERICAN SONG SPARROW, *MELOSPIZA MELODIA*[3]

[3] Swarnali Louha, David A. Ray, Kevin Winker, and Travis C. Glenn, 2020, G3 (Bethesda), 10(4):1159-1166, DOI: 10.1534/g3.119.400929, Reprinted here with permission of publisher.

4.1 ABSTRACT

The song sparrow, *Melospiza melodia*, is one of the most widely distributed species of songbirds found in North America. It has been used in a wide range of behavioral and ecological studies. This species' pronounced morphological and behavioral diversity across populations makes it a favorable candidate in several areas of biomedical research. We have generated a high-quality *de novo* genome assembly of *M. melodia* using Illumina short read sequences from genomic and *in vitro* proximity-ligation libraries. The assembled genome is 978.3 Mb, with a physical coverage of 24.9×, N50 scaffold size of 5.6 Mb and N50 contig size of 31.7 Kb. Our genome assembly is highly complete, with 87.5% full-length genes present out of a set of 4,915 universal single-copy orthologs present in most avian genomes. We annotated our genome assembly and constructed 15,086 gene models, a majority of which have high homology to related birds, *Taeniopygia guttata* and *Junco hyemalis*. In total, 83% of the annotated genes are assigned with putative functions. Furthermore, only ~7% of the genome is found to be repetitive; these regions and other non-coding functional regions are also identified. The high-quality *M. melodia* genome assembly and annotations we report will serve as a valuable resource for facilitating studies on genome structure and evolution that can contribute to biomedical research and serve as a reference in population genomic and comparative genomic studies of closely related species.

4.2 INTRODUCTION

The oscine passerines (Order Passeriformes) are songbirds having specialized vocal learning capabilities (Liu et al. 2013). Many species of songbirds have been widely used by neuroscientists to study the processes underlying memory and learning and social interactions (Doupe and Kuhl 1999, White 2010). The song sparrow (*Melospiza melodia*) is one of the most morphologically diverse songbirds found in North America, with 26 recognized subspecies (Pruett et al. 2008). It has been recognized as a model vertebrate species for field studies of birds and has been the subject of extensive research integrating behavioral and ecological studies over the last 70 years (Arcese et al. 2002). The species is widespread across North America, occupying diverse ecosystems and exhibiting pronounced phenotypic variation in plumage color, seasonal migration and sedentariness, body size, and bill size (Arcese et al. 2002, Pruett & Winker 2010, Greenberg et al. 2012).

Though several species of songbirds have been sequenced and studied (Warren et al. 2010, Jarvis et al. 2014), few offer the plethora of biomedical research potential presented by the song sparrow. This species might serve as a model system in areas such as hepatic lipogenesis (through phenotypic variation in seasonal fat deposition for migration; Gosler 1996, Schubert et al. 2007), craniofacial development (through variation in bill size and shape; Brugmann et al. 2010, Powder et al. 2012), and variations in body size (Sutter et al. 2007, Lango Allen et al. 2010). The latter is a polygenic trait, and elucidation of the underlying gene network affecting different metabolic pathways can help clarify several biological phenomena, including human diseases. Other areas of interest are differences in neural growth and song-center brain development among different song sparrow populations and potential applications in brain neurogenesis (NIH 2001), and also the regeneration of “hair” cells in the song sparrow auditory system and potential therapies useful in hearing loss (Hawkins et al. 2003, Hawkins & Lovett 2004). Given its significant biomedical potential and experimental tractability in the field and aviary, the song sparrow will continue to be used for answering research questions related to mechanisms causing variation in behavior, morphology, and demographics across populations (Arcese et al. 2002, Nietlisbach et al. 2015).

Prior work on song sparrows in Alaska has shown how the song sparrow population in the Aleutian Archipelago is thought to have colonized from the mainland since the last glacial maximum and undergone a series of population bottlenecks to give rise to a naturally inbred population with large body size (Pruett and Winker 2005). The lower genetic variability in this naturally inbred population makes song sparrows from the Aleutian islands a favorable resource for generating a reference genome assembly, because lower levels of polymorphism between both copies of a diploid genome can improve assembly quality. Previous work has also been done on the song sparrow transcriptome, developing genomic markers to screen at population levels (Srivastava et al. 2012). A high-quality genome assembly of *M. melodia* furthers the development of genomic markers to screen loci associated with phenotypic traits of interest. An ever-growing number of songbirds have sequenced genomes, but relatively few have been published so far, including the American crow (*Corvus brachyrhynchos*), golden-collared manakin (*Manacus vitellinus*; Jarvis et al. 2014), Zebra finch (*Taeniopygia guttata*; Warren et al. 2010), medium ground finch (*Geospiza fortis*; Parker et al. 2012) and the dark-eyed junco (*Junco hyemalis*; Friis et al. 2018). In this study, we provide the genome assembly of *Melospiza melodia*, a member of the family Passerellidae. This genome assembly will serve as a reference genome for this species as well as facilitating genomic and phylogenetic comparisons among songbirds and other taxa.

Our high-quality draft genome assembly of *M. melodia* was created by combining both traditional Illumina paired-end libraries and a *de novo* proximity-ligation Chicago library. The Chicago library method together with Dovetail Genomics' HiRise software pipeline is designed to significantly reduce gaps in alignment arising from repetitive elements in the genome (Putnam et al. 2016) and increases assembly contiguity. The draft genome was annotated using transcribed RNA and protein sequences from *M. melodia* and related songbird species, *Junco hyemalis* and *Taeniopygia guttata*. Genomic features of interest other than coding sequences, such as microsatellites, repeat elements, transposable elements, and non-coding RNA, were also annotated and the genome assembly was evaluated for quality by comparing it to related avian species.

4.3 MATERIALS AND METHODS

4.3.1 Library preparation and *de novo* shotgun assembly

The *de novo* assembly of the song sparrow genome was constructed using Illumina paired end libraries. A blood sample from a single male song sparrow was obtained from the wild in the Aleutian Islands of Alaska (Coordinates: 52.8275 / 173.206) on 16 Sep 2003 and archived as a voucher specimen at the University of Alaska Museum (http://arctos.database.museum/guid/UAM:Bird:31500). We chose a male because females are the heterogametic sex in birds and sex chromosomes are known to have highly repetitive DNA content. This together with the selection of an individual from a population known to have lower genetic variation can improve the quality of our assembled genome, without changing the genome structurally. Whole blood was preserved during specimen preparation and shipped overnight in lysis buffer to UGA, where PCI extraction of DNA was performed. We sheared the genomic DNA using a Covaris S2 (Covaris, Woburn, MA, USA) targeting a 600bp average fragment size. The sheared DNA was end-repaired, adenylated, and ligated to TruSeq LT adapters using a TruSeq DNA PCR-Free Library Preparation Kit (Illumina, San Diego, CA, USA). We purified the ligation reaction using a Qiaquick Gel Extraction Kit (Qiagen, Venlo, The Netherlands) from a 2% agarose gel. We sequenced the library on an Illumina HiSeq 2500 at the HudsonAlpha Institute for Biotechnology (Huntsville, AL, USA) to obtain paired-end (PE) ~100 bp reads. The sequence data consisted of 276 million read pairs sequenced from a total of 41.3 Gbp of paired-end libraries (~49× sequencing coverage). Reads were trimmed for quality, sequencing adapters, and mate pair adapters using Trimmomatic (Bolger et al. 2014). The reads were assembled at Dovetail Genomics (Santa Cruz, CA, USA) using Meraculous 2.0.4 (Chapman et al. 2011) with a *k-mer* size of 29. This yielded a 972.4 Mbp assembly with a contig N50 of 22.5 Kbp and a scaffold N50 of 33 Kbp.

4.3.2 Chicago library preparation and scaffolding the draft genome

To improve the *de novo* assembly, a Chicago library was prepared at Dovetail Genomics using previously described methods (Putnam et al. 2016). In brief, about 500 ng of high-molecular-weight genomic DNA (mean fragment length = 50 kbp) was used for chromatin reconstitution *in vitro* and fixed with formaldehyde. Fixed chromatin was digested with *Dpn*II, the 5' overhangs filled in with biotinylated nucleotides, and free blunt ends were ligated together. After ligation, crosslinks were reversed and DNA was purified from protein. Purified DNA was treated to remove biotin that was not internal to ligated fragments. Next, DNA was sheared to ~350 bp mean fragment size and sequencing libraries were generated using NEBNext Ultra enzymes (New England Biolabs, Ipswich, MA, USA) and Illumina-compatible adapters. Biotin-containing fragments were isolated using streptavidin beads before PCR enrichment of the library. The Chicago library was sequenced on an Illumina HiSeq 2500 to produce 47 million 150 bp paired end reads (1-50 kb pairs).

Dovetail Genomics' HiRise scaffolding software pipeline (Putnam et al. 2016) was used to map the shotgun and Chicago library sequences to the draft *de novo* assembly using a modified SNAP read mapper (http://snap.cs.berkeley.edu). The separations of Chicago read pairs mapped within draft scaffolds were analyzed by HiRise to produce a likelihood model for genomic distance between read pairs, and the model was used to identify and break putative misjoins, to score prospective joins, and make joins above a threshold. After scaffolding, shotgun sequences were used to close gaps between contigs.

4.3.3 Identification of microsatellites and transposable elements

Transposable elements (TEs) in the song sparrow genome were identified using a combination of *de novo* and homology-based TE identification methods, in addition to a manual curation step (Platt et al. 2016). First, we used RepeatModeler v1.0.11 (Smit and Hubley 2008-2015) with default parameters (File S1) to generate a custom repeat library consisting of 672 consensus repeat sequences. RepeatModeler uses two *de novo* repeat identification programs, RECON v1.08 (Bao and Eddy 2002) and RepeatScout v1.0.6 (Price et al. 2005), for identifying repetitive elements from sequence data. To ensure accurate and complete representation of putative TEs, the RepeatModeler derived consensus sequences were filtered for size (>100 bp), and then subjected to iterative homology-based searches against the genome, followed by manual curation (Platt et al. 2016). The final set of manually curated TEs was queried against CENSOR (Kohany et al. 2006) and TEclass (Abrusan et al. 2009) for classification. TEs not identifiable in CENSOR were also searched against the NCBI nucleotide and protein databases using BLASTN and BLASTX respectively. Finally, a custom repeat library consisting of 900 repeat elements (File S24) comprising song sparrow-specific TEs and existing repeats in other related avian species was used to screen for repeats in the song sparrow genome assembly with RepeatMasker v4.0.9.

Microsatellites in the song sparrow genome were identified and described with GMATA v2.01 (Wang et al. 2016) with sequence motifs ranging in length from 2-20 bp, and each motif repeated at least 5 times (File S2).

4.3.4 *De novo* gene annotation and function prediction

Genes were predicted in the song sparrow genome with the MAKER v2.31.9 genome annotation pipeline (Campbell et al. 2014). A custom repeat library of 900 repeat sequences (File S24) consisting of TEs identified in the song sparrow genome and other existing avian repeat elements was used to soft mask the genome. Transcriptome evidence sets for MAKER included the assembled song sparrow transcriptome (Srivastava et al. 2012) and Trinity (v2.4.0) mRNA-seq assemblies from multiple tissues of *Junco hyemalis* (Peterson et al. 2012, NCBI BioProject Accession: PRJNA256328). Protein evidence sets used by MAKER included annotated proteins for song sparrow, *Junco hyemalis*, and *Taeniopygia guttata* from the NCBI Protein database. The MAKER pipeline consisted of the following steps: 1) Transcriptomic and protein evidence sets were used to make initial evidence-based annotations with MAKER; 2) the initial annotations were used to train two *ab initio* gene predicters: Augustus (Stanke et al. 2006), which was trained once, and SNAP (Korf 2004), which was iteratively trained twice; and 3) the trained gene prediction tools SNAP and Augustus were used to generate the final set of gene annotations (File S3-S8).

Functional annotations of the predicted genes were obtained by making homology-based searches with BLASTP against the Uniprot/Swiss-Prot protein database (Pundir et al. 2016, File S9). InterProScan v5.29 (Zdobnov and Apweiler 2001) was used to find protein domains associated with the genes. The putative functions and protein domains were added to the gene annotations using scripts provided with MAKER (File S9).

To quantitatively assess the completeness of the song sparrow genome assembly and annotated gene set, we ran BUSCO (Benchmarking Universal Single-Copy Orthologs) v3.0.2 (Waterhouse et al. 2017) with 4,915 single-copy orthologous genes in the Aves lineage group (Aves_odb9; https://busco.ezlab.org/), using "chicken" as the Augustus reference species (File S10). The 4,915 orthologous genes are present in at least 90% of the 40 species included within the Aves lineage group, and thus are likely to be found in the genome of related species. Additionally, we used the JupiterPlot pipeline (https://github.com/JustinChu/JupiterPlot) to visually compare the zebra finch (*T. guttata*) genome assembly (Warren et al. 2010) to our assembly in a Circos plot, using the largest scaffolds making up 85% of our genome assembly, and all scaffolds greater than 100 kbp in the Zebra finch genome (File S11). We also used the JupiterPlot pipeline to compare our assembly to the genome assemblies of the collared flycatcher (*Ficedulla albicollis*), great tit (*Parus major*) and house sparrow (*Passer domesticus*). These birds were selected for comparison because they have highly complete genomes, and are often used for comparative genomic studies in birds.

4.3.5 Non-coding RNA prediction

Transfer RNAs (tRNAs) were predicted in the song sparrow genome with tRNAscan-SE v2.0 (Lowe and Chan 2016, File S12). A training set comprising eukaryotic tRNAs was used to train the covariance models employed by tRNAscan-SE, and tRNAs were searched against the genome with Infernal v1.1.2 (Nawrocki 2014). tRNAscan-SE also provides functional classification of tRNAs based on a comparative analysis using a suite of isotype-specific tRNA covariance models. A random sample of 10 predicted tRNAs were selected and searched against the tRNA databases GtRNAdb (Chan and Lowe 2016) and tRNAdb (Jühling et al.2009).

Identification of miRNAs (microRNAs), snoRNAs (small nucleolar RNAs), snRNAs (small nuclear RNAs), rRNAs (ribosomal RNAs), and lncRNAs (long non-coding RNAs) was achieved by using a homology-based prediction method. Structural homologs to eukaryotic ncRNA covariance models from the Rfam database v14.1 (Gardner et al. 2009) were searched against the song sparrow genome using Infernal's (v1.1.2) "cmscan" program (File S13). All low-scoring overlapping hits and hits with an E-value greater than 10^{-5} were discarded, and the remaining ncRNAs were grouped into different classes.

Lastly, we compared the predicted classes of different ncRNAs in the song sparrow genome to those reported in the genomes of related birds, *Taeniopygia guttata* and *Ficedula albicollis* (collared flycatcher).

4.4 DATA AVAILABILITY

Raw reads have been deposited in the NCBI Sequence Read Archive (SRR10491484 and SRR10451714 for the Meraculous assembly, and SRR10424475 for the Chicago HiRise assembly). The *M. melodia* Chicago HiRise genome sequence (Mmel_1.0), and annotations are available in GenBank under the accession RZID00000000 (NCBI BioProject accession: PRJNA511035). Supplemental File S1 contains submission script for RepeatModeler. Supplemental File S2 contains primary configuration file used to run GMATA (default_cfg.txt). Supplemental File S3 contains submission script for MAKER. Supplemental File S4 contains MAKER executable file (maker_exe.ctl). Supplemental File S5 contains specifications for downstream filtering of BLAST and Exonerate alignments (maker_bopts.ctl). Supplemental File

S6 contains primary configuration of MAKER specific options (maker_opts.ctl). Supplemental File S7 contains scripts for training SNAP. Supplemental File S8 contains scripts for training Augustus. Supplemental File S9 contains scripts for running BLASTP and InterProScan for functional annotation of predicted genes; and scripts for adding the functional annotations to gene annotation files. Supplemental File S10 contains submission script for BUSCO. Supplemental File S11 contains submission scripts for JupiterPlot pipeline. Supplemental File S12 contains submission script for tRNAscan-SE. Supplemental File S13 contains submission script for Infernal. Supplemental File S14 contains classification of predicted transposable elements. Supplemental File S15 contains annotation of microsatellites with their genomic locations. Supplemental File S16 contains percentage of different microsatellites present in the genome. Supplemental File S17 contains frequency of occurrence of microsatellites in each scaffold of the genome. Supplemental File S18 contains the distribution of the length of microsatellites. Supplemental File S19 contains predicted function of annotated genes by BLASTP. Supplemental File S20 contains prediction of protein domains, GO annotations and pathway annotations of predicted genes by InterProScan. Supplemental File S21 contains sequence and structure of tRNAs identified in the song sparrow genome. Supplemental File S22 contains classification of predicted tRNAs. Supplemental File S23 contains classification of different ncRNAs predicted in the genome with Infernal. Supplemental File S24 contains custom repeat library used to screen for repeats in the song sparrow genome. Supplemental Table S1 contains genome sizes of birds related to *M. melodia*. Supplemental Figure S1 contains the distribution of the percentage of annotated genes with their corresponding AED scores. Supplemental Figure S2 contains the distribution of the top base-pair composition of microsatellite motifs in the *M. melodia* genome. Supplemental Figure S3 contains comparison of the *M. melodia* genome assembly with genome assemblies of related birds. Supplemental material available at figshare: https://doi.org/10.25387/g3.11676441.

4.5 RESULTS AND DISCUSSION

4.5.1 Assembly

We produced the *de novo* genome assembly of song sparrow, with a total length of 978.3 Mb, using a Chicago library and the HiRise assembly pipeline. The N50 scaffold size was 5.6 Mb and contig size was 31.7 Kb. This assembly showed significant improvement over the initial shotgun assembly, with a 169-fold increase in scaffold N50 and a 60-fold increase in scaffold N90 (Table 4.1). These increases in scaffold size were also accompanied by an increase in assembly contiguity, with the total number of scaffolds decreasing from 74,832 to 13,785 (Fig 4.1, Table 4.1).

4.5.2 Microsatellites and Transposable Elements

In total, 88 as yet unnamed TEs were identified in the song sparrow genome. Fifty-five of these did not have any significant matches in CENSOR (Kohany et al. 2006) and are considered novel (File S14). A TE was considered to have a significant match to a known element in CENSOR only when it had a length of at least 80 bp and 80% identity to the known element over 80% of its length, the 80-80-80 rule (Wicker et al. 2007). The predicted TEs were classified into DNA transposons and retrotransposons (i.e. LINEs, LTRs, and SINEs) using CENSOR and TEclass (File S14). Approximately 7.4% of the genome comprises repeats with the majority of that consisting of TEs (~ 48%). Among the different TEs, LTRs (~ 40%) and LINEs (~ 49%) were found to be most abundant (Table 4.2). The song sparrow genome assembly was found to be less repetitive when compared to sequenced genomes of related songbirds, primarily due to the lower content of LTRs and LINEs than other songbirds (Fig 4.2).

Overall, 112,419 microsatellites with motifs ranging in size from 2-20 bp were found in the song sparrow genome (File S15 contains all microsatellites with their genomic locations). The majority of the microsatellites were made up of 2-, 3-, 4-, and 5-mers, with 2-mers making up about 71% of all microsatellites identified (Fig 4.3, File S16). The distribution of the top base-pair composition of microsatellite motifs present in the genome is shown in Fig S2. The frequency of occurrence of microsatellites in every scaffold and a distribution of their lengths are provided in Files S17 and S18, respectively.

4.5.3 Gene annotation and function prediction

The MAKER genome annotation pipeline predicted 15,086 genes and 139 pseudogenes in the song sparrow genome, fewer than *T. guttata, F. albicollis*, and *M. vitellinus*, but higher than *G. fortis* (Table 4.3). The average gene length, exon length, intron length, and the total number of exons and introns predicted are also less compared to closely related species (Table 4.3). Of the 15,086 predicted genes, 12,541 genes were assigned putative functions with BLASTP (File S19). InterProScan assigned functional domains to 11,298 (74.9%) predicted genes (File S20). A total of 7,010 genes obtained GO annotations. Pathway annotations were assigned to 2,716 genes.

Annotated genes were assigned annotation edit distance (AED) scores with values ranging from 0 to 1. AED is a distance metric score that signifies how closely gene models match transcript and protein evidence. Gene models with AED scores closer to 0 have better alignment with the evidence provided in the MAKER pipeline. A distribution of the percentage of genes with their corresponding AED scores shows close similarity of the annotated genes with the transcript and protein evidence provided in the MAKER pipeline (Fig S1).

The song sparrow genome assembly contained 4,318 complete universal single-copy orthologs (BUSCOs; 87.9%) from a total of 4,915 BUSCO groups searched. Among all complete BUSCOs, 99.4% were present as single-copy genes and 0.6% were duplicated. About 7.4% (356) of the orthologous gene models were partially recovered, and 4.9% (241) had no significant matches. The incomplete and missing gene models could either be partially present or missing, or could indicate genes that are too divergent or have very complex structures, making their prediction difficult. Incomplete and missing gene models could also suggest problems associated with the genome assembly and gene annotation. The results from the BUSCO analysis are in agreement with the Circos plot (Fig 4.4), in which few scaffolds in the *T. guttata* genome assembly are not represented in our assembly and very few inconsistent arrangements of scaffolds exist between the two genome assemblies. Comparison of our assembly to *F. albicollis*, *P. major*, and *P. domesticus* genome assemblies showed many more inconsistencies in the arrangements of scaffolds between the genomes of these birds and *M. melodia* (Fig S3) than between *T. guttata* and *M. melodia*.

4.5.4 Non-coding RNA prediction and identification

A total of 267 tRNAs were detected in the song sparrow genome by tRNAscan-SE (see File S21 for sequence and structure of tRNAs), out of which 129 were found coding for the standard twenty amino acids. The predicted output from tRNAscan-SE (File S22) contained 114 tRNAs with low Infernal as well as Isotype scores; these were characterized as pseudogenes lacking tRNA-like secondary structures (Lowe and Chan 2016). Two tRNAs had undetermined isotypes and 22 were chimeric, with mismatch isotypes. Chimeric tRNAs contain point mutations in their anticodon sequence, rendering different predicted isotypes than those predicted by structure-specific tRNAscan-SE covariance models. Among all predicted tRNAs, 11 contained introns within their sequences. No suppressor tRNAs and tRNAs coding for selenocysteine were predicted. The subset of 10 randomly selected tRNAs was also predicted in many other species in both GtRNAdb and tRNAdb databases.

Infernal searches predicted a total of 364 ncRNAs in the song sparrow genome, comprising 166 miRNAs, 8 rRNAs, 154 snoRNAs, 16 snRNAs, and 20 lncRNAs (File S23). Compared to the genomes of related avian species (*T. guttata* and *F. albicollis*), the song sparrow genome has the highest number of predicted tRNAs, but fewer other ncRNAs (Table 4.4).

4.6 CONCLUSION

The Chicago and shotgun sequencing libraries along with the HiRise assembly software enabled accurate and highly contiguous *de novo* assembly of the song sparrow genome. The genome assembly is 978.3 Mb, with 48 scaffolds (L50) making up half the genome size. A previous estimate of genome size of *M. melodia* from densitometry analysis provided a C-value of 1.43 pg (1,398.54 Mb) (Andrews et al. 2009). Our own *k-mer* based estimate of genome size from paired reads in the shotgun and Chicago libraries using Kmergenie v1.7044 (Chikhi and Medvedev 2014) yielded an estimated size of 1,127.25 Mb. Both these genome size estimates and the genome sizes of related birds (Table S1) are slightly higher than our genome assembly (978.3 Mb). Our small assembly size may be attributed to the compression of repetitive regions, which is generally observed in assemblies generated from short-read sequencing data. This is also consistent with the fact that our genome contains fewer repeats when compared to related songbirds (Fig 4.2). Although short reads limit our ability to characterize the total number of repeats within long tandem arrays, we have been able to characterize vast majority of repeats, resolving them into LINEs, SINEs, LTRs, and DNA retrotransposons (Fig 4.2, Table 4.2).

Our genome is highly complete, with 87.5% full-length genes present out of 4,915 universal orthologous genes in avian species. A large set of genes (15,086) with known homology to related birds was annotated in our study. A majority of these genes (83%) were assigned with putative functions. The improved scaffold lengths and gene model annotations will facilitate studies to identify genes responsible for multiple phenotypic traits of interest. Additionally, longer scaffolds in the Chicago HiRise assembly will help detect regions under selection, including SNPs and structural variants such as insertions/deletions or copy number variations which are potentially responsible for the phenotypic diversity observed in this species.

Though we report fewer miRNAs, snRNAs, snoRNAs, rRNAs, and lncRNAs in this genome than in related songbirds, we have high confidence in the predicted ncRNAs we report because we used conservative cutoffs to reduce false positives. Pending the availability of long-read data, this genome assembly provides an excellent reference for a range of genetic, ecological, functional, and comparative genomic studies in song sparrows and other songbirds.

4.7 ACKNOWLEDGMENTS

This research was supported in part by an Institutional Development Award (IDeA) from the National Institute of General Medical Sciences of the NIH (P20GM103395). The content is solely the responsibility of the authors and does not necessarily reflect the official views of the NIH. We thank David Sonneborn and Jack Withrow for their roles in making a vouchered specimen available for our work. We also thank Troy Kieran for assistance with laboratory work and sequencing logistics, Roger Nilsen and the staff of Georgia Genomics and Bioinformatics Core for constructing the genomic library, and the staff of Dovetail Genomics for help in preparing and processing the Chicago library and HiRise assemblies. The high performance computing cluster at Georgia Advanced Computing Resource Center (GACRC) at the University of Georgia provided computational infrastructure and technical support throughout the work.

4.8 REFERENCES

Abrusan G, Grundmann N, DeMester L, Makalowski W. 2009. TEclass: a tool for automated classification of unknown eukaryotic transposable elements. Bioinformatics 25:1329-1330.

Andrews CB, Mackenzie SA, Gregory TR. 2009. Genome size and wing parameters in passerine birds. Proc Biol Sci 276:55-61.

Arcese P, Sogge MK, Marr AB, Patten MA. 2002. Song Sparrow (*Melospiza melodia*), version 2.0. *In* Poole AF, Gill FB (ed), The Birds of North America, Cornell Lab of Ornithology, Ithaca, NY, USA.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114-2120.

Brugmann SA, Powder KE, Young NM, Goodnough LH, Hahn SM, James AW, Helms JA, Lovett M. 2010. Comparative gene expression analysis of avian embryonic facial structures reveals new candidates for human craniofacial disorders. Hum Mol Genet 19:920-930.

Campbell MS, Holt C, Moore B, Yandell M. 2014. Genome annotation and curation using MAKER and MAKER-P. Curr Protoc Bioinformatics 48:4.11.1-39.

Chan PP, Lowe TM. 2016. GtRNAdb 2.0: An expanded database of transfer RNA genes identified in complete and draft genomes. Nucleic Acids Res 44:D184-D189.

Chapman JA, Ho I, Sunkara S, Luo S, Schroth GP, Rokhsar DS. 2011. Meraculous: de novo genome assembly with short paired-end reads. PLoS One 6:e23501.

Chikhi R, Medvedev P. 2014. Informed and automated k-mer size selection for genome assembly. Bioinformatics 30:31-37.

Doupe AJ, Kuhl PK. 1999. Birdsong and Human Speech: Common Themes and Mechanisms. Annu Rev Neurosci 22:567-631.

Friis G, Fandos G, Zellmer AJ, McCormack JE, Faircloth BC, Milá B. 2018. Genome-wide signals of drift and local adaptation during rapid lineage divergence in a songbird. Mol Ecol 27:5137-5153.

Gardner PP, Daub J, Tate J, Moore BL, Osuch IH, Griffiths-Jones S, Finn RD, Nawrocki EP, Kolbe DL, Eddy SR, Bateman A. 2011. Rfam: Wikipedia, clans and the 'decimal' release. Nucleic Acids Res 39:D141-D145.

Gosler AG. 1996. Environmental and social determinants of winter fat storage in the Great Tit *Parus major*. J Anim Ecol 65:1-17.

Greenberg R, Cadena V, Danner RM, Tattersall GJ. 2012. Heat Loss May Explain Bill Size Differences between Birds Occupying Different Habitats. PLoS ONE 7:e40933.

Hawkins RD, Bashiardes S, Helms CA, Hu L, Saccone NL, Warchol ME, Lovett M. 2003. Gene expression differences in quiescent versus regenerating hair cells of avian sensory epithelia: implications for human hearing and balance disorders. Hum Mol Genet 12:1261-1272.

Hawkins RD, Lovett M. 2004. The developmental genetics of auditory hair cells. Hum Mol Genet 13:R289-R296.

Jarvis ED, Mirarab S, Aberer AJ, Li B, Houde P, et al. 2014. Whole-genome analyses resolve early branches in the tree of life of modern birds. Science 346:1320-1331.

Jühling F, Mörl M, Hartmann RK, Sprinzl M, Stadler PF, Pütz J. 2009. tRNAdb 2009: Compilation of tRNA sequences and tRNA genes. Nucleic Acids Res 37:D159-D162.

Kohany O, Gentles AJ, Hankus L, Jurka J. 2006. Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. BMC Bioinformatics 7:474.

Korf I. 2004. Gene finding in novel genomes. BMC Bioinformatics 5:59.

Lango AH, Estrada K, Lettre G, Berndt SI, Weedon MN, et al. 2010. Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature 467:832-838.

Liu WC, Wada K, Jarvis ED, Nottebohm F. 2013. Rudimentary substrates for vocal learning in a suboscine. Nat Commun 4:2082.

Lowe TM, Chan PP. 2016. tRNAscan-SE On-line: Search and Contextual Analysis of Transfer RNA Genes. Nucleic Acids Res 44:W54-W57.

Nawrocki EP. 2014. Annotating functional RNAs in genomes using Infernal. Methods Mol Biol 1097:163-197.

Nietlisbach P, Camenisch G, Bucher T, Slate J, Keller LF, Postma E. 2015. A microsatellite-based linkage map for song sparrows (*Melospiza melodia*). Mol Ecol Resour 15:1486-1496.

NIH. 2001. What we learned from songbirds: The adult brain can generate new nerve cells. NIH Publication No. 01-4602.

Parker P, Li B, Li H, Wang J. 2012. The genome of Darwin's Finch (*Geospiza fortis*). GigaScience. Available at http://dx.doi.org/10.5524/100040.

Peterson MP, Whittaker DJ, Ambreth S, Sureshchandra S, Buechlein A, Podicheti R, Choi JH, Lai Z, Mockatis K, Colbourne J, Tang H, Ketterson ED. 2012. De novo transcriptome sequencing in a songbird, the dark-eyed junco (*Junco hyemalis*): genomic tools for an ecological model system. BMC Genomics 13:305.

Platt RN 2nd, Blanco-Berdugo L, Ray DA. 2016. Accurate Transposable Element Annotation Is Vital When Analyzing New Genome Assemblies. Genome Biol Evol 8:403-10.

Powder KE, Ku YC, Brugmann SA, Veile RA, Renaud NA, Helms JA, Lovett M. 2012. A cross-species analysis of microRNAs in the developing avian face. PLoS One 7:e35111.

Price AL, Jones NC, Pevzner PA. 2005. De novo identification of repeat families in large genomes. Bioinformatics 21 Suppl 1:i351-i358.

Pruett CL, Arcese P, Chan YL, Wilson AG, Patten MA, Keller LF, Winker K. 2008. Concordant and discordant signals between genetic data and described subspecies of Pacific Coast Song Sparrows. The Condor 110:359-364.

Pruett CL, Winker K. 2005. Northwestern Song Sparrow populations show genetic effects of sequential colonization. Mol Ecol 14:1421-1434.

Pruett CL, Winker K. 2010. Alaska Song Sparrows (*Melospiza melodia*) demonstrate that genetic marker and method of analysis matter in subspecies assessments. Ornithological Monographs 67:162-171.

Pundir S, Martin MJ, O'Donovan C, The UniProt Consortium. 2016. UniProt Tools. Curr Protoc Bioinformatics 53:1.29.1-1.29.15.

Putnam NH, O'Connell BL, Stites JC, Rice BJ, Blanchette M, Calef R, Troll CJ, Fields A, Hartley PD, Sugnet CW, Haussler D, Rokhsar DS, Green RE. 2016. Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. Genome Res 26:342-350.

Schubert KA, Mennill DJ, Ramsay SM, Otter KA, Boag PT, Ratcliffe LM. 2007. Variation in social rank acquisition influences lifetime reproductive success in black-capped chickadees. Biol J Linn Soc 90:85-95.

Smit AFA, Hubley R. 2008-2015 RepeatModeler Open-1.0.11. Available at http://www.repeatmasker.org.

Srivastava A, Winker K, Shaw TI, Jones KL, Glenn TC. 2012. Transcriptome analysis of a North American songbird, *Melospiza melodia*. DNA Res 19:325-333.

Stanke M, Tzvetkova A, Morgenstern B. 2006. AUGUSTUS at EGASP: using EST, protein and genomic alignments for improved gene prediction in the human genome. Genome Biol 7 Suppl 1:S11.1-8.

Sutter NB, Bustamante CD, Chase K, Gray MM, Zhao K, Zhu L, Padhukasahasram B, Karlins E, Davis S, Jones PG, Quignon P, Johnson GS, Parker HG, Fretwell N, Mosher DS, Lawler DF, Satyaraj E, Nordborg M, Lark KG, Wayne RK, Ostrander EA. 2007. A single IGF1 allele is a major determinant of small size in dogs. Science 316:112-115.

Wang X, Wang L. 2016. GMATA: An Integrated Software Package for Genome-Scale SSR Mining, Marker Development and Viewing. Front Plant Sci 7:1350.

Warren WC, Clayton DF, Ellegren H, Arnold AP, Hillier LW, et al. 2010. The genome of a songbird. Nature 464:757-62.

Waterhouse RM, Seppey M, Simão FA, Manni M, Ioannidis P, Klioutchnikov G, Kriventseva EV, Zdobnov EM. 2017. BUSCO applications from quality assessments to gene prediction and phylogenomics. Mol Biol Evol 35:543-548.

White SA. 2010. Genes and vocal learning. Brain Lang 115:21-28.

Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P, Morgante M, Panaud O, Paux E, SanMiguel P, Schulman AH. 2007. A unified classification system for eukaryotic transposable elements. Nat Rev Genet 8:973-982.

Bao Z, Eddy SR. 2002. Automated de Novo Identification of Repeat Sequence Families in Sequenced Genomes. Genome Res 12:1269-1276.

Zdobnov EM, Apweiler R. 2001. InterProScan-an integration platform for the signature-recognition methods in InterPro. Bioinformatics 17:847-848.

Zhang G, Li C, Li Q, Li B, Larkin DM, et al. 2014. Comparative genomics reveals insights into avian genome evolution and adaptation. Science 346:1311-1320.

Tables:

Table 4.1: A comparison of assembly quality statistics from the initial shotgun sequencing assembled by Meraculous and the final HiRise assembly.

	Meraculous Assembly	**Chicago HiRise Assembly**
Total length	972.4 Mb	978.3 Mb
Scaffold N50	33 kb	5.58 Mb
Scaffold N90	5 kb	303 kb
Scaffold L50	7,552 scaffolds	48 scaffolds
Scaffold L90	35,731 scaffolds	324 scaffolds
Longest scaffold	366,149	26,942,064
Number of scaffolds	74,832	13,785
Number of scaffolds > 1kb	74,806	13,768
Contig N50	22.5 kb	31.7 kb
Number of gaps	53,577	95,490
Percent of genome in gaps	1.427%	1.847%
Number of N’s per 100 kbp	1427.15	1847.03
GC content	41.07%	41.08%

Table 4.2: Number and percentage of repeats in the *M. melodia* genome assembly.

Classification	**Number of copies**	**Percentage of assembly**
LINEs	104,032	3.01
LTRs	85,276	2.83
SINEs	6,695	0.08
DNA Transposons	13,521	0.21
Unclassified	4,884	0.12
Total transposable elements	**214,408**	**6.25**
Satellites	569	0.00
Low complexity repeats	38,561	0.20
Microsatellites	192,996	0.90
Total	**446,534**	**7.35**

Table 4.3: Characteristics of genes predicted in the *M. melodia* genome compared to *Taeniopygia guttata* (zebra finch), *Ficedulla albicollis* (collared flycatcher), *Manacus vitellinus* (golden-collared manakin) and *Geospiza fortis* (medium ground finch).

	M. melodia	*T. guttata*[1]	*F. albicollis*[2]	*M. vitellinus*[3]	*G. fortis*[4]
Number of genes	15,086	17,561	16,763	18,976	14,388
Mean gene length (bp)	14,457	26,458	31,394	27,847	30,164
Mean CDS length (bp)	1,325	1,677	1,942	1,929	1,766
Number of exons	131,940	171,767	189,043	190,390	164,721
Mean exon length (bp)	153	225	253	264	195
Mean number of exons/gene	8.67	10.25	12.22	11.51	11.41
Number of introns	116,724	153,909	171,236	171,089	149,563
Mean intron length (bp)	1,695	2,930	3,257	3,294	2,813

[1] https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Taeniopygia_guttata/103/
[2] https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Ficedula_albicollis/101/
[3] https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Manacus_vitellinus/102/
[4] https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Geospiza_fortis/101/

Table 4.4: Number of ncRNAs predicted in the *Melospiza melodia* genome compared to *Taeniopygia guttata* (zebra finch) and *Ficedulla albicollis* (collared flycatcher).

	M. melodia	*T. guttata*[1,2]	*F. albicollis*[1,3]
tRNA	267	184	179
miRNA	166	302	510
snRNA	16	44	32
snoRNA	154	241	199
rRNA	8	100	22
lncRNA	20	908	1473

[1] http://useast.ensembl.org/info/data/ftp/index.html
[2] https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Taeniopygia_guttata/103/
[3] https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Ficedula_albicollis/101/

Figures:

Figure 4.1: Comparison of assembly contiguity.

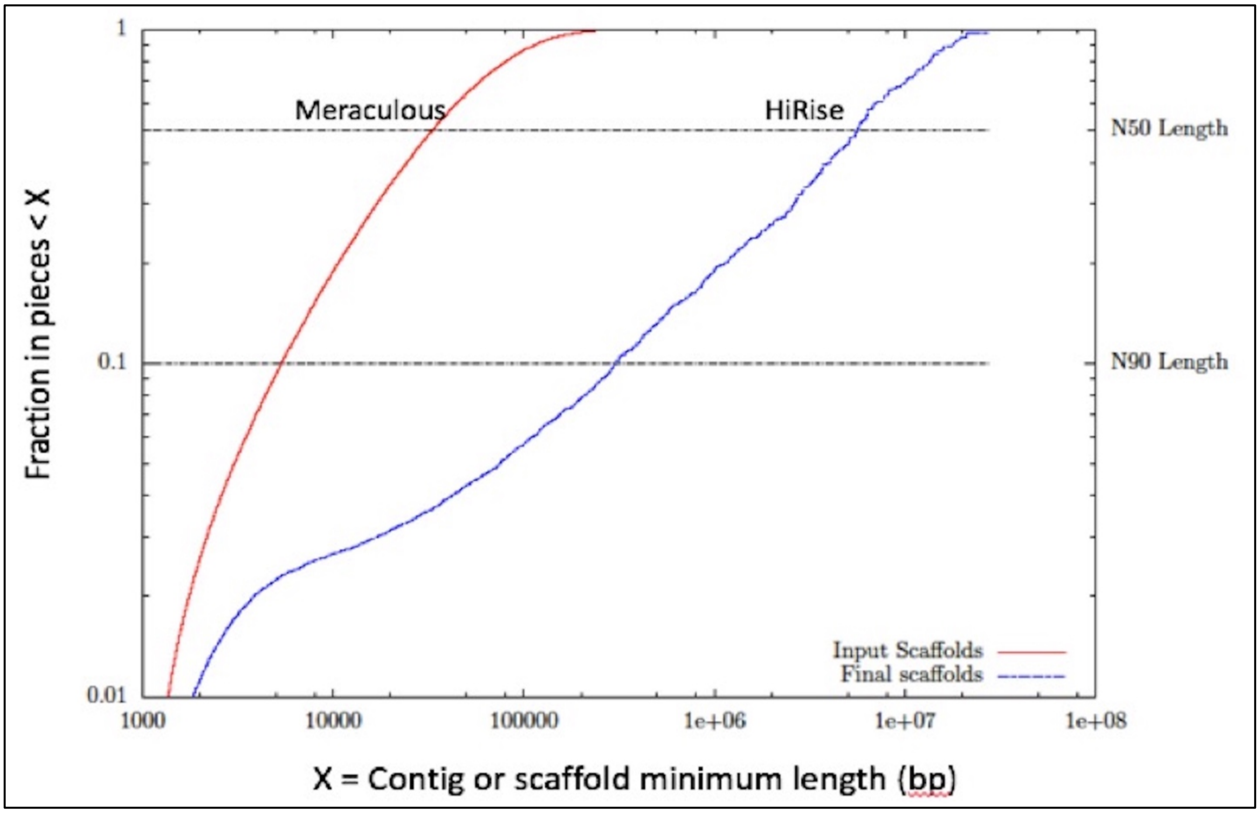


Figure 4.2: Comparison of percentages of transposable elements (TEs) among related songbird genome assemblies. * Data from: Zhang *et al*. 2014 Science. 346: 1311-1320.

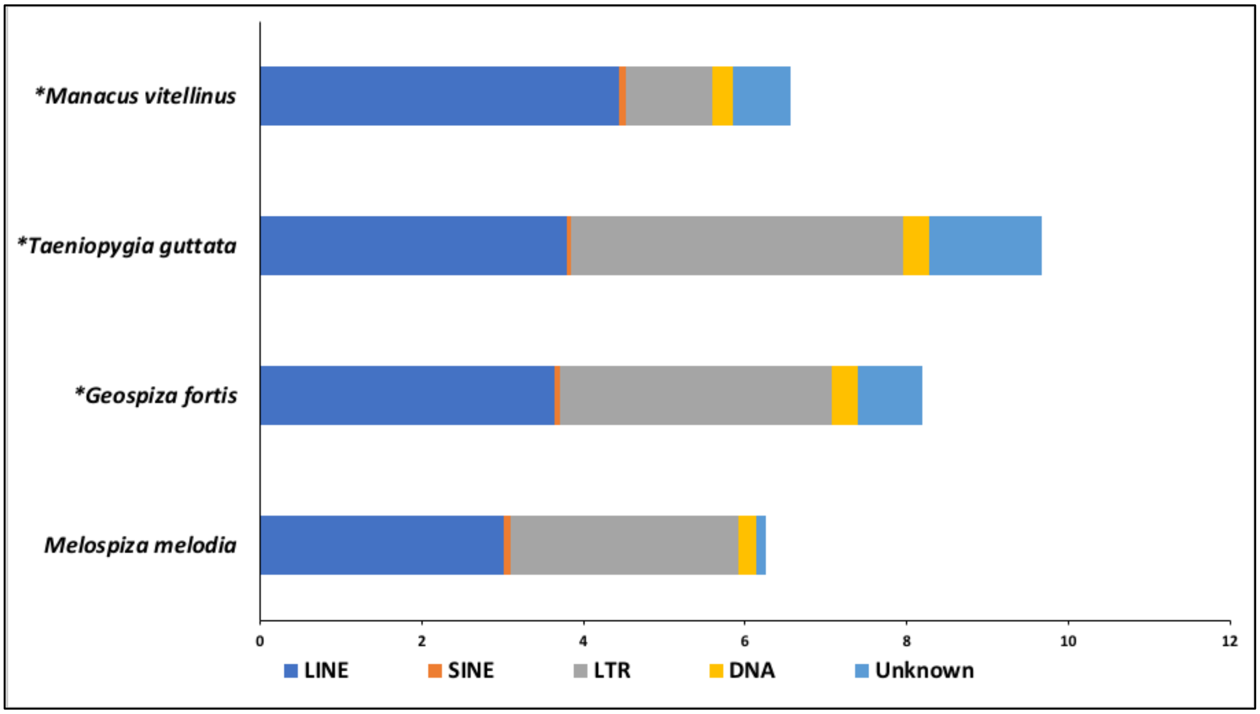


Figure 4.3: Abundance of microsatellite repeat motif size classes in the *M. melodia* genome assembly (details are given in Supplemental File S16).

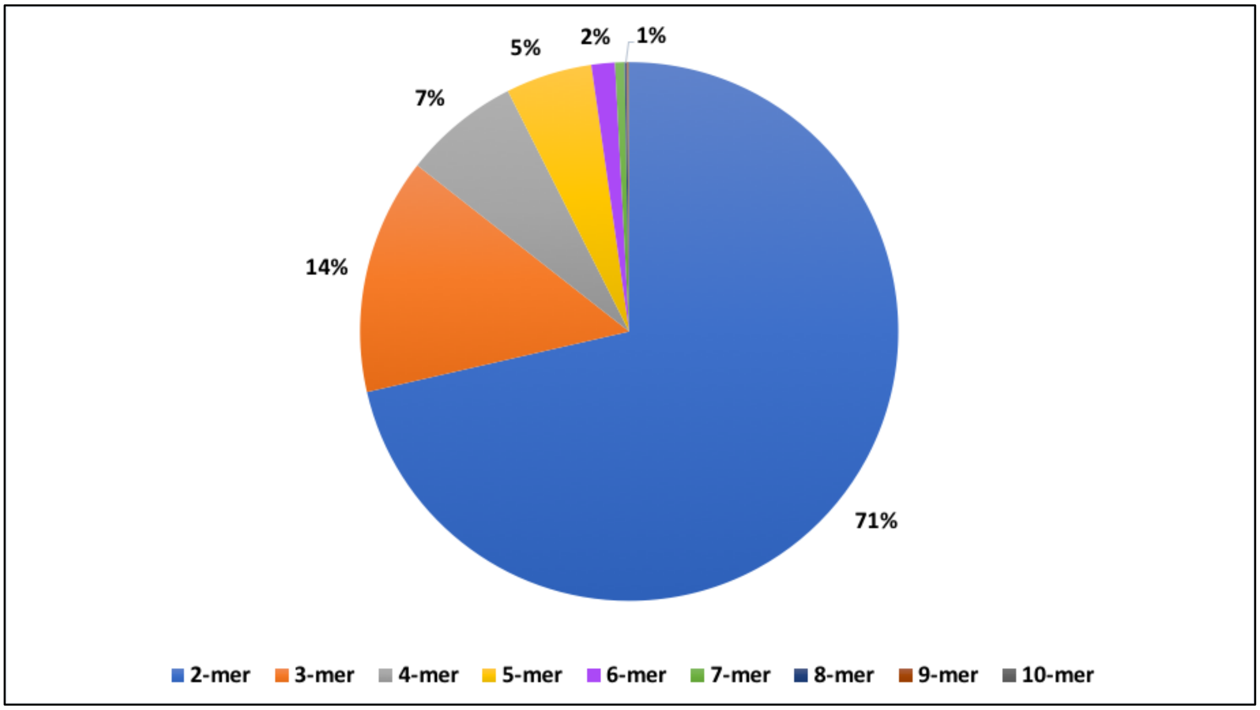
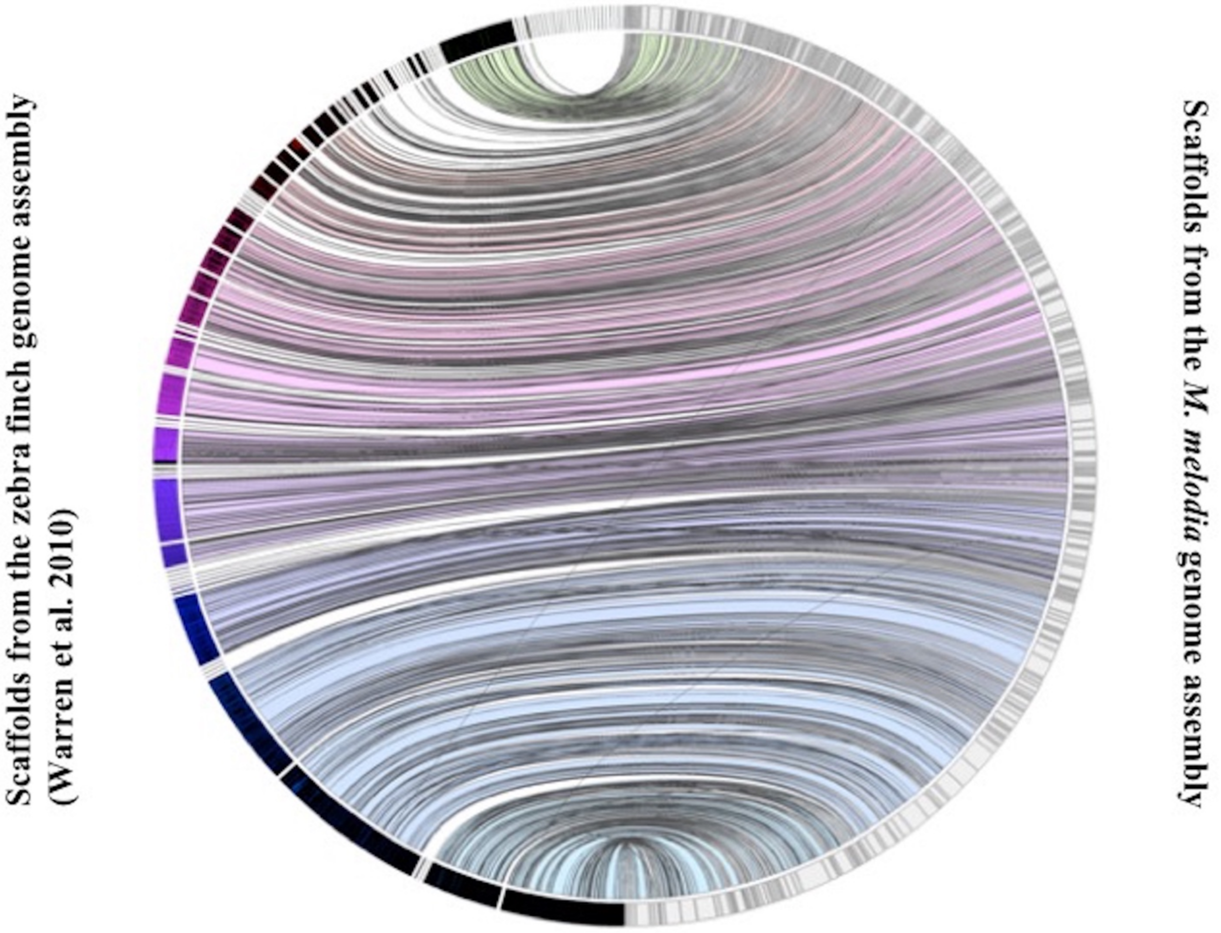


Figure 4.4: Jupiter plot correlating zebra finch and song sparrow genome assemblies, considering scaffolds greater than 100 kbp in the reference zebra finch genome and the largest scaffolds representing 85% of the song sparrow genome.



4.9 SUPPLEMENTAL MATERIAL

Supplemental material is available at https://doi.org/10.25387/g3.11676441

CHAPTER 5

CONCLUSION AND FUTURE WORK

5.1 GENOME CHARACTERIZATION OF *LISTERIA MONOCYTOGENES*

In chapter 2, I developed an open source tool, Haplo-ST, that can perform wgMLST-based characterization of isolates of *Listeria monocytogenes* from whole-genome sequencing datasets. This tool helped us classify two groups of *L. monocytogenes* isolates collected from different ecological sources (outdoor environment and poultry processing plants) into distinct sequence types and clonal complexes, and evaluate the phylogenetic relationships between members of each group. Additionally, genetic differentiation studies performed on the wgMLST profiles of isolates obtained from both groups provided insights into loci potentially contributing towards increased adaptability and persistence of *L. monocytogenes* in poultry processing facilities. Haplo-ST can not only be used for characterization of *L. monocytogenes* isolates, but can also be extended to classify and evaluate phylogenetic relationships between isolates of other haloid organisms, simply by installation of an organism-specific gene database and making minor modifications to the script that automates the pipeline. Further developments to this tool could involve enabling automated allele curation for new alleles not present in the gene database and construction of a module that can create minimum spanning trees from isolate subtype data. Further, because long read sequencing projects have become increasingly popular in recent years, future work could also involve development of applications that can assemble alleles from data generated by third-generation sequencers like PacBio and Oxford Nanopore. Use of long read sequencing technologies for allele assembly can also enable assembly and characterization of differences in regions of the genome other than protein-coding genes. Thus, the power of fully assembled genomes can be exploited for bacterial classification.

Secondly, our tool was used to assess patterns of linkage disequilibrium among protein-coding genes in the genome of *L. monocytogenes* (chapter 3). Our analysis revealed presence of strong linkage disequilibrium among majority of genes in the genome of this species. This analysis also helped us detect genes which were less significantly associated to other genes in the genome and we considered these genes to be potential “hot spots” for horizontal gene transfer. Future extensions of this project can involve application of this approach to other bacterial species such as *Salmonella enterica* and Mycobacterium tuberculosis, both of which have a highly clonal genetic structures (Liu et al. 2006, Didelot et al. 2011, Yar et al. 2018). This will not only give us insights into whether the patterns of linkage disequilibrium obtained in *L. monocytogenes* are similar to other highly clonal bacteria or are unique to it, but also help reveal the processes contributing to the diversification and evolution of these microbes.

5.2 GENOME ANALYSIS OF *MELOSPIZA MELODIA*

The primary objective of our genome analysis project (chapter 4) was to produce a high-quality genome assembly of the song sparrow, *Melospiza melodia*, and obtain annotations of protein-coding genes, repeats and different classes of non-coding RNAs. We have achieved all these goals by using our genome analysis pipeline. We believe that the *M. melodia* genome assembly and associated annotations will serve as valuable resources for studying the genome structure and evolution in this species and also contribute towards population and comparative genomic studies in closely related avian species.

The song sparrow is one of the most polytypic bird species found in North America, with 26 recognized sub-species that exhibit great morphological variation across their range (Patten and Pruett 2009). The largest members of this species reside in the Aleutian archipelago and has almost three times the body size of the subspecies found in California. The reference genome produced in this project can be used for understanding the processes that affect the physiology of this species such that it attains weight, resulting in the gigantism phenotype. Our resources can also be used for comparing the genomes of different sub-species of song sparrows and examining the patterns of genomic change that result in their divergence and evolution. The molecular approach by which we plan to achieve these goals is to sequence a large number of individuals of this species using low coverage whole-genome sequencing techniques, RADseq (Davey and Blaxter 2010), sequence capture etc. and study the genetic variation responsible for producing the highly diverse phenotypic characteristics. Variation between different sub-species of song sparrows can also be studied using other molecular markers like microsatellites or SNPs observed from sequencing data.

5.3 REFERENCES

Davey JW, Blaxter ML. 2010. RADSeq: next-generation population genetics. Brief Funct Genomics 9:416-423.

Didelot X, Bowden R, Street T, Golubchik T, Spencer C, McVean G, Sangal V, Anjum MF, Achtman M, Falush D, Donnelly P. 2011.Recombination and population structure in *Salmonella enterica*. PLoS Genet 7:e1002191.

Liu X, Gutacker MM, Musser JM, Fu YX. 2006.Evidence for recombination in *Mycobacterium tuberculosis*. J Bacteriol 188:8169-77.

Patten M, Pruett C. 2009. The Song Sparrow*, Melospiza melodia*, as a ring species: patterns of geographic variation, a revision of subspecies, and implications for speciation. Systematics and Biodiversity 7:33-62.

Yar AM, Zaman G, Hussain A, Changhui Y, Rasul A, Hussain A, Bo Z, Bokhari H, Ibrahim M. 2018. Comparative Genome Analysis of 2 *Mycobacterium* tuberculosis Strains from Pakistan: Insights Globally Into Drug Resistance, Virulence, and Niche Adaptation. Evol Bioinform Online 14:1176934318790252.