ABSTRACT

Title of Dissertation: **POPULATION AND GENETIC DIVERSITY ANALYSIS OF *LISTERIA MONOCYTOGENES* IN SELECT FOODS AND FOOD PROCESSING ENVIRONMENTS**

Hee Jin Kwon, Doctor of Philosophy, 2023

Dissertation directed by: Professor, Jianghong Meng
Department of Nutrition and Food Science

Listeria monocytogenes, a Gram-positive bacterium, is a foodborne pathogen that causes listeriosis in humans. *L. monocytogenes* can persist in various environmental conditions, including food-relevant conditions such as high salinity, refrigerated temperatures, and low moisture contents. Contaminated food products, including dairy products, deli meats, fresh produce, and soft cheeses, are the primary transmission vehicles for *L. monocytogenes*. The complex and dynamic population structure of *L. monocytogenes* complicates control efforts, particularly due to certain strains that may possess increased resistance to stress conditions and/or enhanced virulence. The advent of whole genome sequencing has facilitated comprehensive genomic analyses of *L. monocytogenes*, enabling a comprehensive understanding of its adaptation and survival characteristics over time and across various geographic locations. Understanding the population and genetic diversity of *L. monocytogenes* is crucial for the development of effective control measures, as it helps infer the spread and transmission pathways of *L. monocytogenes* through the integration of spatial-temporal

factors. Furthermore, these analyses provide insights into the evolutionary relationships among *L. monocytogenes* strains. This dissertation aimed to investigate the population diversity of *L. monocytogenes* in various food sources and food processing facilities, utilizing the whole genome sequencing technology. The findings contribute valuable insights into the genetic diversity and population structure of *L. monocytogenes,* thereby aiding the understanding of the risk associated with *L. monocytogenes* contamination and the development of effective control measures to ensure food safety.

POPULATION AND GENETIC DIVERSITY ANALYSIS OF *LISTERIA MONOCYTOGENES* IN SELECT FOODS AND FOOD PROCESSING ENVIRONMENTS

by

Hee Jin Kwon

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2023

Advisory Committee:
Professor Jianghong Meng, Chair
Dr. Yi Chen
Professor Abani Pradhan
Professor Qin Wang
Professor Yanjin Zhang

Acknowledgments

The journey to complete my dissertation at the University of Maryland would not be as pleasant and rewarding without the help and support of many people. Foremost, I would like to express my sincere appreciation to my advisors, Dr. Jianghong Meng and Dr. Yi Chen, for their unwavering support and encouragement throughout my Ph.D. program. Additionally, I am immensely grateful to the members of my committee, Dr. Abani Pradhan, Dr. Qin Wang, and Dr. Yanjin Zhang, for their invaluable expertise, time, and guidance during this journey. I would also like to extend my gratitude to all my colleagues at the Center for Food Safety and Applied Nutrition, Food and Drug Administration, whose support and camaraderie have been instrumental in my progress. Lastly, I would like to express my deepest thanks to my family and friends for their abundant love, unwavering support, and continuous encouragement throughout this entire time.

Table of Contents

List of Tables

List of Figures

Figure IV-4. Phylogenetic tree based on the distance matrix generated using CFSAN SNP pipeline.

List of Abbreviations

AOAC	The Association of Official Agricultural Chemists
BEAST	Bayesian Evolutionary Analysis Sampling Trees
BLAST	Basic Local Alignment Search Tool
CCs	Clonal Complexes
CDC	The Centers for Disease Control and Prevention.
CFSAN	The Center for Food Safety and Applied Nutrition
ESS	Effective Sample Size
FDA	Food and Drug Administration
HKY	The Hasegawa–Kishino–Yano
HPD	Highest Posterior Density
LGI	*Listeria* genomic island
LIPI	*Listeria* pathogenicity islands
MCC	Maximum Clade Credibility
MCMC	The Markov chain Monte Carlo
MGE	Mobile Genetic Elements
MLST	Multi-locus Sequencing Typing
tMRCA	Time of Most Recent Common Ancestor
MST	Minimum Spanning Tree
NCBI	The National Center for Biotechnology Information
ONT	Oxford Nanopore Technologies
PCR	Polymerase Chain Reaction
PFGE	Pulse Field Gel Electrophoresis

PHASTER	PHAge Search Tool Enhanced Release
PMSC	Premature Stop Codons
QUAST	Quality Assessment Tool for Genome Assemblies
RTE	Ready To Eat
SNPs	Single Nucleotide Polymorphisms
SRA	Sequence Read Archive
SSI	Stress Survival Islet
STs	Seqeunce Types
WGS	Whole Genome Sequencing

Chapter I: Introduction

Listeria monocytogenes is a Gram-positive, facultatively anaerobic, non-spore-forming bacterium responsible for causing listeriosis, an opportunistic foodborne disease that primarily affects vulnerable populations such as pregnant women, the elderly, and immunocompromised individuals **[1]**. The symptoms of listeriosis can vary from mild gastrointestinal issues in the general population to severe conditions such as meningitis, encephalitis, mother-to-fetus infections, and septicemia, resulting in high rates of hospitalization and case fatality **[2]**. *L. monocytogenes* is widely distributed in nature, including soil, water, vegetation, and animals, due to its ability to survive across a wide range of temperatures, pH values, moisture content, and salinity levels **[3]**. This ability enables *L. monocytogenes* to thrive in various ecological niches associated with food production, such as food contact surfaces, equipment, and biofilms, where the bacteria can persist for extended periods and have multiple potential transmission routes to food products **[4]**. Consequently, the presence of *L. monocytogenes* remains a significant concern for food safety due to its ability to survive and persist, and its potential to cause significant foodborne illness.

Heterogenicity of *L. monocytogenes*

L. monocytogenes is a genetically heterogeneous species with a highly clonal population structure, where each strain possesses distinct genetic profiles, phenotypic characteristics, and virulence potential **[5]**. *L. monocytogenes* can be categorized into four major evolutionary lineages, each of which includes several serotypes **[6, 7]**. Lineage I consists of strains with serotypes 1/2b, 3b, 4b, 4d, 4e, and 7, while lineage II contains strains with serotypes 1/2a, 1/2c, 3a, and 3c. Lineage III consists of strains with serotypes 4b, 1/2a, 4a, and 4c, and lineage IV contains strains with serotypes 4a and 4c **[8]**. More than 95% of *L. monocytogenes* strains associated with food and clinical cases belong to lineages I and II, with serotypes 1/2a, 1/2b, and 4b being the most prevalent **[6]**. The evolutionary lineages can be further divided into multiple clonal complexes (CCs) consisting of sequence types (STs) defined by multi-locus sequencing typing (MLST) **[5]**. Analysis of genetic diversity in *L. monocytogenes* has revealed that certain clones were strongly associated with clinical samples, indicating that they are more likely to cause severe infections in humans and designated as hypervirulent clones **[9]**. Other clones are considered hypovirulent, which are commonly found in food and environmental samples but are less invasive **[9]**. Hypervirulent clones of *L. monocytogenes* often carry major genes associated with its pathogenicity, such as *Listeria* pathogenicity island (LIPI)-3 and LIPI-4 **[9]**, as well as genes contributing to enhanced stress tolerance. LIPI-3 encodes the bacteriocin listeriolysin S enhancing hemolytic and cytotoxic reactions **[10].** LIPI-4 is responsible for the infection of *L. monocytogenes* in neural and placental tropisms **[9]**. Another major virulence factor is the intact *inlA* gene, which encodes the inlA protein required

for invading host epithelial cells **[11]**. *L. monocytogenes* isolates carrying premature stop codons (PMSCs) in *inlA,* which causes a truncation of inlA, would manifest attenuated virulence of *L. monocytogenes* **[12]**. This genetic diversity of *L. monocytogenes* can arise from various mechanisms, including horizontal gene transfer, genetic recombination, and the accumulation of spontaneous mutations **[13]**. These molecular evolution processes of *L. monocytogenes* can result in the emergence of strains that exhibit enhanced abilities to resist stress conditions and possess increased virulence potential. Thus, understanding the genetic diversity within *L. monocytogenes* population is critical for comprehending its adaptive capabilities and the factors contributing to its persistence and pathogenicity.

Phenotypic and molecular subtyping methods for *L. monocytogenes*

Subtyping of foodborne pathogens plays a critical role in outbreak investigations as it enables the identification and tracking of bacterial strains responsible for foodborne illness **[14]**. Different subtyping approaches are used for the characterization of *L. monocytogenes*, involving phenotypic examination or genotypic analysis of isolates **[15]**. The serological analysis is a phenotypic subtyping method for *L. monocytogenes*, targeting the species-specific surface proteins on the bacterial cells, such as somatic (O) and flagellar (H) antigens. The serotypes of *L. monocytogenes* strains are determined based on the combination of O and H antigen subtypes, with 15 O antigens subtypes and four H antigens subtypes **[16]**. Among the 13 serotypes identified for *L. monocytogenes*, the four main serotypes found in food and clinical samples are 1/2a, 1/2b, 1/2c, and 4b **[17, 18]**. However, serotyping with

traditional agglutination methods is time-consuming and demands good technical practice, making it challenging to build a standardized database. Polymerase chain reaction (PCR) has enabled molecular-based subtyping by detecting specific genetic markers or patterns in bacterial strains **[19]**. The PCR-based approach for *L. monocytogenes* serotyping was proposed as an alternative to the traditional agglutination, which involves the amplification of four genetic markers, such as *lmo0737, lmo1118,* ORF2819, and ORF2110 **[20]**. Based on the presence or absence of these targeted genetic markers, strains can be categorized into four major serogroups: serogroups IIa (1/2a and 3a), IIb (1/2b, 3b and 7), IIc (1/2c and 3c) and IVb (4b, 4d and 4e) **[20]**. The utilization of PCR in serogroup identification allows a rapid classification of *L. monocytogenes* with standardized discriminatory power.

Another molecular subtyping approach for *L. monocytogenes* is called MLST, which involves analyzing the nucleotide sequences of seven housekeeping genes (i.e., *abcZ*, *bglA*, *cat*, *dapE*, *dat, ldh,* and *lhkA*). MLST is a well-established technique that provides unambiguous data though a public database **[21, 22]**. These genes are relatively conserved within the species but exhibit enough sequence diversity to enable discrimination between different strains. Moreover, MLST facilitates the construction of phylogenetic trees and genetic analyses to understand the population structure and evolutionary linkage between strains **[23]**.

Pulse-field gel electrophoresis (PFGE) has been considered the gold standard molecular subtyping method for investigating outbreaks of *L. monocytogenes* due to its highly discriminatory power and reproducibility **[14]**. This technique utilizes rare-cutting restriction enzymes (such as *Asc*I and *Apa*I) to digest the genomic DNA **[24]**.

Subsequently, the resulting DNA fragments are separated on a gel, creating distinct PFGE patterns that serve as DNA fingerprints for a particular strain **[25]**. PFGE plays an important role in an outbreak investigation by comparing PFGE patterns with a national database of PFGE patterns maintained by the Centers for Disease Control and Prevention (CDC) **[24]**.

The advent of next-generation sequencing technologies has significantly reduced the cost of DNA sequencing, making whole genome sequencing (WGS) an accessible and valuable tool for the investigation of foodborne outbreaks **[26]**. WGS involves sequencing the complete genome of a pathogen, providing unparalleled discriminatory capability and enabling high-resolution and comprehensive genomic characterization. Thus, it has become the ultimate tool for molecular subtyping of bacterial isolates **[26]**. Moreover, WGS enables the identification of single nucleotide polymorphisms (SNPs) and core-genome MLST, which targets over 1,000 genes, providing the highest level of discrimination in strain levels. The WGS-based approaches allow for the construction of phylogenetic trees that illustrate the evolutionary relationships of closely related isolates, integrating spatial-temporal factors **[27, 28]**.

Listeriosis outbreaks in the United States

There are estimated 1,600 individuals affected by listeriosis every year, and about 260 of those cases result in fatalities in the U.S. **[29].** The first listeriosis outbreak recognized in the U.S. was associated with pasteurized milk in 1983 **[30]**. Since then, multistate listeriosis outbreaks have been repeatedly linked to certain food

vehicles, such as dairy products, ready-to-eat (RTE) deli meat products, and vegetables. These products pose a high risk for listeriosis due to the lack of additional heating procedures before consumption and their extended storage periods at refrigerated temperatures. *L. monocytogenes* can survive and even proliferate during the refrigeration storage in these products, thereby increasing the potential risk of causing infections.

In recent years, there have been multistate listeriosis outbreaks recognized and linked to novel or atypical food vehicles, partially due to advancements in outbreak surveillance and the use of WGS approaches for traceback efforts **[31]**. Many of these atypical food vehicles were previously regarded as low-risk foods for *L. monocytogenes* growth due to unfavorable conditions. However, in the past two decades, seven listeriosis outbreaks have been reported, associated with atypical food products such as ice cream, boiled eggs, frozen vegetables, and intact fruit commodities. Despite these occurrences, not or rarely known regarding *L. monocytogenes* behaviors in these atypical food vehicles, implying major knowledge gaps in our understanding.

Project overview

Given the frequent isolation of *L. monocytogenes* in foods and food production environments, it is critical to actively monitor the prevalence of this pathogen and investigate population structure in select foods and food processing facilities. The specific objectives for each chapter are outlined below:

1) To identify and characterize the mobile genetic elements (MGEs) associated with persistent *L. monocytogenes* from meat and poultry processing facilities by taking advantage of the long-read sequencing technology and to investigate the evolution of CC6 strains persistent in meat and poultry processing facilities and those involved in major meat and poultry-associated outbreaks worldwide. The frequent recovery of *L. monocytogenes* isolates belonging to CC6 showing outbreak PFGE patterns from meat and poultry processing facilities suggested the potential persistence of strains over several years. It was hypothesized that the MGEs are unique genetic regions, acquired through horizontal gene transfer, thatcan be utilized as genetic markers to track the persistence and spread of *L. monocytogenes*.
2) To assess the genetic diversity of *L. monocytogenes* isolates recovered from whole fresh avocados. An unexpectedly high prevalence of *L. monocytogenes* was observed during the 2014 FDA surveillance, suggesting that the whole fresh avocados can serve as a food vehicle for *L. monocytogenes*. The genetic diversity was investigated to identify *L. monocytogenes* clones that can be unique to avocado products directly out of production, those from retail and those from different geographic locations The whole genome sequencing data

were also used to track the possible transmission events between avocados grown and distributed in different geographic locations.

3) To determine the prevalence and genetic diversity of *L. monocytogenes* isolated from ice cream production facilities in the U.S. In 2016, *L. monocytogenes* isolates were detected during the national surveillance in the ice cream production facilities. Genetic diversity was assessed to investigate genetic characteristics, identify unique and/or dominant clones and determine possible transmission patterns of *L. monocytogenes.* Furthermore, *L. monocytogenes* strains collected during the listeriosis outbreak investigation linked to the contaminated ice cream products in 2015 were included for a comparative analysis to infer possible transmission among facilities and food products.

.

Chapter II: Characterization of mobile genetic elements for tracking persistent *L. monocytogenes* recovered from ready-to-eat meat and poultry processing facilities.

Abstract

Recently developed nanopore sequencing technologies offer a unique opportunity to rapidly close the genome and to identify complete sequences of mobile genetic elements (MGEs). In this chapter, 17 isolates of *L. monocytogenes* clonal complex 6 (CC6) from seven ready-to-eat meat or poultry processing facilities, not known to be associated with outbreaks, were shotgun sequenced, and among them, five isolates were further subjected to long-read sequencing. Additionally, 26 genomes of *L. monocytogenes* CC6 isolates associated with three listeriosis outbreaks in the U.S. and South Africa were obtained from the National Center for Biotechnology Information (NCBI) database and analyzed to evaluate if MGEs may be used as a high-resolution genetic marker for identifying and sourcing the origin of *L. monocytogenes*. The analyses identified four *comK* prophages in 11 non-outbreak isolates from four facilities and three *comK* prophages in 20 isolates associated with two outbreaks that occurred in the U.S. In addition, three different plasmids were identified among 10 non-outbreak isolates and 14 outbreak isolates. Further characterization of the MGEs determined that they were highly conserved among the isolates from the same incident. Different prophages from different facilities or

outbreaks had genetic variations, possibly due to horizontal gene transfer. In addition, phylogenetic analysis showed that isolates from the same facility or the same outbreak are always closely clustered. The time of the most recent common ancestor of the *L. monocytogenes* CC6 isolates was estimated to be in March 1816, with the average nucleotide substitution rate of 3.1×10^{7} substitutions per site per year. This study showed that complete MGE sequences provide a good signal to determine the genetic relatedness of *L. monocytogenes* isolates, to identify persistence or repeated contamination that occurred within the food processing environment, and to study the evolutionary history among closely related isolates.

Introduction

The presence of *L. monocytogenes* in RTE meat products has been a significant food safety concern, as these products have a prolonged shelf life and are typically consumed without any additional heating or cooking procedures **[32]**. The most likely source of contamination is associated with persistent *L. monocytogenes* clones in food processing facilities, and these clones can survive and grow in RTE meat products under refrigerated storage and high salinity conditions, increasing the risk of listeriosis infections among consumers. Since 1990, multiple incidents of *L. monocytogenes* contamination in RTE meat products have been reported, establishing RTE meat products as a typical food vehicle for *L. monocytogenes* **[3]**. In 1998, a novel genotype of *L. monocytogenes* serotype 4b, which was later designated as CC6 strains, was first recognized, implying a multistate listeriosis outbreak. The outbreak was attributed to the contamination of frankfurters and deli meats, which led to 108 cases, including 14 deaths in the U.S. **[33]**. Another multistate listeriosis outbreak linked to CC6 strains was reported in 2002, involving contaminated turkey deli products and resulting in 54 cases, including 8 deaths **[34]**. A multi-virulence-locus sequence typing comparative study conducted on the 1998 and 2002 *L. monocytogenes* outbreak isolates revealed a close genetic relationship between two outbreak-associated isolates, suggesting a possible transmission of *L. monocytogenes* between meat processing plants **[35]**. More recently, *L. monocytogenes* CC6 strains were implicated in a major listeriosis outbreak in South Africa from 2017 to 2018, attributed to the contamination of RTE meat products, and it was recognized as the largest listeriosis outbreak worldwide, with a total of 1,060 cases **[36]**. These

incidents highlighted the need for a comprehensive understanding of the contamination, persistence, and transmission of *L. monocytogenes* CC6 strains.

WGS has expanded the scope of molecular surveillance to not only verify the link between patients and food sources but also to identify genetic evolution among isolates within the same or different clones. Further development and widespread use of next-generation sequencing have significantly advanced our understanding of genomic changes during bacterial evolution. Comparative genomics and evolutionary analysis on a small portion of prophage regions adjacent to the phage insertion sites of *L. monocytogenes* CC6 outbreak isolates in 1998 and 2002 revealed that the prophage region may be conserved in isolates associated with the same outbreak, implying a short-term evolution scenario **[37, 38]**; however, these studies did not review the entire prophage regions. A similar phenomenon was observed in isolates from CCs associated with other listeriosis outbreaks that were linked to different types of foods when the entire prophage regions of these isolates were analyzed **[39]**. In addition to the prophage, a plasmid harbored by *L. monocytogenes* CC6 isolates associated with the 1998 listeriosis outbreak was characterized based on PCR amplicon and shotgun sequencing **[40]**. The PCR amplicon sequencing, targeting known resistance genes, such as *bcrABC* or *tmr*, conferring resistance to benzalkonium chloride (BC) and other disinfectants, was used to characterize the plasmid found in the 1998 U.S. outbreak isolates **[41, 42]**. Thus, the diversity and dynamics of MGEs, such as prophages and plasmids, can be essential to understand the adaptation and persistence of *L. monocytogenes* and to identify resident strains in the food processing facility. However, the absence of the complete prophage and

plasmid sequences in these previous studies has hindered the assessment of whether these MGEs also exhibit adequate variations among strains to enable high resolution mapping of source populations.

Gaining insight into the short-term evolution that occurred within the same clone can aid in understanding the ability of *L. monocytogenes* to persist despite environmental stresses. For instance, it was demonstrated that *L. monocytogenes* strains carrying *comK* prophages exhibit increased persistence and rapid adaptation via biofilm formation in the food processing facilities, increasing the potential risk of repeated contamination **[37]**. Plasmids may similarly contain genes encoding important stress resistance features, such as resistance to BC and other disinfectants widely used in food processing facilities **[43]**. The persistent strain(s) could have acquired resistance genes via horizontal gene transfer, which has been shown to occur through the transmission of MGEs **[44]**. Plasmid-borne resistance genes, such as *bcrABC,* which confers resistance to BC, were previously reported on a putative transposon harbored by a plasmid of H7550, a 1998 listeriosis outbreak isolate, suggesting evidence of a role for MGEs contributing to adaptation and persistence of *L. monocytogenes* strains in the meat or poultry processing facilities **[42]**. The gene cassette, *bcrABC*, was also observed in plasmids of multiple isolates from a 2017 outbreak associated with ice cream in Florida, but the cassette was not observed in plasmids of strains from other listeriosis outbreaks analyzed in that study **[39]**. Other plasmid-borne resistance genes, such as *qacA*, *qacC*, *emrE*, and *emrC*, were observed in plasmids of *L. monocytogenes* strains isolated from the retail environment in the U.S. **[45, 46]**.

L. monocytogenes prophages and plasmids are often larger than 30 Kb and contain multiple repetitive regions **[47, 48]**, challenging whole-genome shotgun sequencing based on short sequence reads (≤ 300 bp) to precisely assembly these repeat sequences **[49]**. A *de novo* assembly of shotgun sequencing then produces incomplete and fragmented genomes; thus, it may constrain us from capturing the entire MGEs sequences. Recently developed long-read sequencing generating reads spanning the repetitive sequences enabled us to obtain a complete genome by closing gaps in fragmented assemblies **[49, 50]**. However, compared with short-read sequencing, the assembled genomes obtained by long-read sequencing often showed a relatively high error rate, indicating assemblies based solely on Oxford Nanopore Technologies (ONT) may not accurately represent the true genome sequences **[51]**. To overcome the relative limitations of relying solely on long-read or short-read sequencing technologies, a hybrid assembly pipeline that combines both technologies can be utilized to generate complete genomes **[52]**.

In this chapter, we employed long-read sequencing from ONT to complete the genomes of *L. monocytogenes* CC6 isolates previously recovered up to 22 months apart from RTE meat or poultry processing environments, indicating possible persistence or repeated contamination [7]. Then we identified, characterized, and quantified variations among the complete MGEs based on WGS data of CC6 isolates sequenced in this chapter, as well as previously sequenced isolates associated with listeriosis outbreaks involving RTE meat or poultry products. Lastly, we assessed the evolutionary rate of these CC6 isolates to gain a better understanding of the population structure of this *L. monocytogenes* clone.

Materials and Methods

Isolates and whole-genome sequencing

Seventeen *L. monocytogenes* CC6 isolates were obtained from seven RTE meat or poultry processing facilities (A, B, C, D, E, F, and G) from two months apart to 22 months apart (**Table II-1**) **[37]**. These isolates are not known to be associated with any outbreaks and thus were referred to as non-outbreak isolates hereinafter. Genomic DNA of each isolate was extracted from overnight cultures using QIAcube apparatus (QIAGEN Inc., Valencia, CA, USA) following the manufacturer's manuals for Gram-positive bacteria. The library for Illumina shotgun sequencing was prepared using a Nextera XT Library Preparation Kit (Illumina, Inc., San Diego, CA, USA). The sequencing was performed on an Illumina MiSeq platform with a 500-cycle Illumina MiSeq Reagent Kit v2 (Illumina, Inc., San Diego, CA, USA). Paired-end reads (2 × 250 bp) were trimmed using Trimmomatic v0.36.4 **[53]** with default parameters and de novo assembled using SPAdes v3.12.0 **[54]**. If our analysis of MGEs predictions revealed inconsistencies between the SPAdes assemblies and the complete genome predictions, we utilized SKESA v0.24 **[55]** to assemble an alternative assembly of the raw reads.

After identifying isolates that were anticipated to contain one or more MGEs, we performed long-read sequencing to obtain complete genomes for these isolates. High-quality genomic DNA extraction was performed using Qiagen Genomic Tip 500/G columns (QIAGEN Inc., Valencia, CA, USA) per the manufacturer's instructions. DNA library preparation for a long-read sequencing was followed using a 1D DNA ligation sequencing kit (SQK-LSK109) (Oxford Nanopore Technologies

Table II-1. Seventeen *L. monocytogenes* CC6 isolates recovered from seven meat or poultry processing facilities.

Isolate	Source	State	Facility	Date of isolation	*comK* gene	PHASTER	PlasmidFinder -2.0
OB020621	Food	NC	A	Sep-2002	Disrupted	Identified	No
OB020790	Food	NC	A	Nov-2002	Disrupted	Identified	No
OB030029	Food	IN	B	Jan-2003	Disrupted	Identified	Yes
OB040119	Food	PA	C	Jun-2004	Disrupted	Identified	Yes
OB050272	Food	PA	C	Aug-2005	Disrupted	Identified	Yes
OB050226	Food	PA	D	Jul-2005	Disrupted	Identified	Yes
OB050347	Environmental	PA	D	Oct-2005	Disrupted	Identified	Yes
OB050350	Environmental	PA	D	Oct-2005	Disrupted	Identified	Yes
OB050351	Environmental	PA	D	Oct-2005	Disrupted	Identified	Yes
OB050355	Environmental	PA	D	Oct-2005	Disrupted	Identified	Yes
OB070122	Environmental	PA	D	May-2007	Disrupted	Identified	Yes
OB080183	Food	NY	E	May-2008	Intact	Not identified	Yes
OB080396	Environmental	NJ	F	Aug-2008	Intact	Not identified	No
OB080398	Environmental	NJ	F	Aug-2008	Intact	Not identified	No
OB080487	Environmental	NJ	F	Oct-2008	Intact	Not identified	No
OB080567	Environmental	NJ	F	Dec-2009	Intact	Not identified	No
OB090318	Environmental	NY	G	Aug-2009	Intact	Not identified	No

Inc., Oxford, UK) and DNA concentrations were measured using a Qubit dsDNA HS assay kit (Fisher Scientific Inc., Hampton, NH, USA). The prepared libraries were loaded into an ONT MinION flow cell (R9.4.1) and sequenced on the MinION device. The sequenced reads were base-called in real-time using MinKnow 3.4.8 integrated with Guppy 3.0.7. Long-read data were then assembled with corresponding shotgun data to obtain complete genomes **[56]** using a hybrid assembly pipeline in Unicycler v0.4.8 **[57]**. Briefly, short reads were first *de novo* assembled, and long reads were used to close gaps and build bridges. Final complete sequences were polished multiple times using its short raw reads from Illumina MiSeq as a reference by Pilon v1.23 **[57, 58]**. Genome coverage was calculated based on the total base pairs of initial raw reads divided by the final length of chromosome and plasmid.

Additionally, 26 publicly available *L. monocytogenes* CC6 isolates associated with three listeriosis outbreaks involving contaminated RTE meat or poultry products were selected for comprehensive analyses. A total of 22 shotgun genomes and five complete genomes of the outbreak isolates were retrieved from the NCBI database. These included nine isolates from the 1998 U.S. outbreak, 11 isolates from the 2002 U.S. outbreak, and six isolates from the 2017–2018 South African outbreak (**Table II-2**). Notably, the two environmental isolates associated with the 2002 U.S. outbreak (J1816 and J1815) were obtained from two different facilities, with J1816 (Facility X) likely linked to the outbreak and exhibiting the same PFGE pattern. On the other hand, J1815 did not have the same PFGE profile, and it was isolated from a different facility (Facility Y), but still considered to be associated with the 2002 U.S. listeriosis outbreak in the U.S. **[34]**.

Table II-2. Outbreak associated *L. monocytogenes* CC6 isolates included this chapter.

Isolate	Source	State	Date of isolation	NCBI accession number or SRA
The 1998 U.S. outbreak isolates				
H7355	Clinical	Unknown	Nov-1998	SRR1814362
H7738	Food	OH	Dec-1998	SRR3707884
H7762	Food	MI	Nov-1998	SRR3707885
H7961	Food	OH	Jan-1999	SRR1814399
H7962	Food	OH	Jan-1998	SRR1815438
H7550	Clinical	NY	Oct-1998	SRR1815437
H7557	Food	NY	Nov-1998	SRR3707886
H7596	Food	NY	Sep-1998	SRR1815440
H7969	Clinical	OH	Jan-1998	SRR3707879
The 2002 U.S. outbreak isolates				
J1776	Clinical	NJ	Sep-2002	NC_021839.1
J1816	Environmental	PA	Oct-2002	NC_022047.1 /SRR2544677
J1817	Environmental	PA	Oct-2002	NC_021830.2
J1926	Food	PA	Nov-2002	NC_021840.1
J1703	Clinical	PA	Sep-2002	SRR3707894
J1705	Clinical	PA	Sep-2002	SRR3707893
J1735	Clinical	Unknown	Jan-2002	SRR3707726
J1736	Clinical	PA	Sep-2002	SRR1815439
J1815	Environmental	MI	Jan-2002	SRR3707728
J1925	Food	Unknown	Nov-2001	SRR1814333
J1927	Food	PA	Nov-2002	SRR3707892
The 2017 South African outbreak isolates				
HM00113468	Food	Unknown	Feb-2018	NZ_CP058256
HM00108598	Food	Unknown	Jan-2018	SRR7056256
HM00110618	Environmental	Unknown	Feb-2018	SRR7056247
IG01149260	Clinical	Unknown	Dec-2017	SRR7056255
YA00079283	Clinical	Unknown	Dec-2017	SRR7056251
YA00082404	Food	Unknown	Jan-2018	SRR7056250

SNP-based phylogenetic analysis

The CFSAN SNP Pipeline v1.0.1 **[59]** was used to perform phylogenetic analysis on the whole-genome shotgun raw data of the 22 outbreak-associated isolates and 17 non-outbreak isolates. Paired-end reads were aligned to the complete genome of OB020621 as a reference using Bowtie2 **[60],** and the variant sites were identified using VarScan2 **[61]**. The reference genome OB020621 was chosen because it did not carry plasmids. The pipeline excludes high-density variant regions that contain at least three variant sites in any 1000 bp span to avoid recombination events or an assembly error **[62]**. A SNP matrix was generated and used to construct a maximum likelihood tree with 1000 bootstraps values in MEGA X **[63]**. The SNP alignment of 39 isolates generated by CFSAN SNP Pipeline was further used for a tip-dated analysis as described below.

Identification of MGEs: plasmids and chromosome-borne prophages

Based on previous studies suggesting that many of these CC6 non-outbreak isolates contained *comK* prophages **[37]**, we identified the *comK* gene (609 bp) from shotgun genomes to predict the possible presence of a *comK* prophage in a shotgun genome. The complete *comK* gene was identified in the one isolate from Facility E, all isolates from Facility F, and the one isolate from Facility G. Therefore, one isolate from four facilities (A, B, C, and D) was first chosen to be sequenced by ONT MinION for genome closure. Putative prophages were predicted using PHASTER via a web portal **[64]** from all complete genomes and shotgun genomes. All "intact" prophages were subjected to further analysis to confirm the presence and completeness of prophage. "Incomplete" predictions found in shotgun genomes were

considered only when the putative prophages were found at the end of a contig. A "questionable" prophage prediction of a 10.7 Kb length was observed in all isolates but not considered since it was previously described as monocin **[40]**. We then reviewed the genes in putative prophages, and also compared the locations of putative prophages and the *comK* gene. If a putative prophage was determined to be a *comK* prophage, we used the disrupted *comK* gene to modify its start and end locations. Then the complete sequences of identified *comK* prophages were compared using Artemis Comparison Tool v18.1.0 **[65]**.

For the isolates of the 1998 U.S. outbreak from which no complete genomes were available, we performed PHASTER analysis on shotgun genomes and also aligned the *comK* gene to determine the possible presence of *comK* prophage. For the isolates of the 2002 U.S. outbreak and the 2017–2018 South African outbreak, we performed PHASTER analysis on complete genomes and shotgun genomes. If a putative *comK* prophage was identified, we modified its start and end locations using the disrupted *comK* gene. After PHASTER analysis of individual genome, we then used BLAST: (i) to compare predicted prophages from different isolates, especially those predicted as "intact" prophages in the complete genomes and in the middle of shotgun contigs, (ii) to identify the prophage that was shared by all isolates or any prophage unique to specific isolate(s), and (iii) to establish if the prophages predicted are present in other complete or shotgun genomes.

PlasmidFinder-2.0 [66] designed for Gram-positive bacteria was used to predict potential plasmids, and also used BLAST (https://blast.ncbi.nlm.nih.gov) to compare shotgun genomes with all *L. monocytogenes* plasmids published in the NCBI

database to identify any plasmid contigs and to calculate the total length of each plasmid. We predicted fragments of plasmids from shotgun genomes in isolates from Facilities B, C, D, and E. There was no plasmid predicted from the isolates of Facilities A, F, and G. Thus, OB080183 from Facility E was additionally chosen to be sequenced by ONT MinION.

A discrete closed extrachromosomal sequence that was produced by the Unicycler hybrid assembler indicated the presence of a plasmid in that MinION-sequenced genome (Table 2). Subsequently, BLAST was used to search for those MinION-closed plasmids within the shotgun genomes that were assembled without nanopore sequencing data, with significant matches used to determine the presence/absence of these plasmids in shotgun genomes. The plasmids were then aligned with selected complete *L. monocytogenes* plasmids published in the NCBI using Mauve alignment v1.1.1 **[67]** to determine if the plasmids were novel or previously reported from other *L. monocytogenes* strains.

In the NCBI database, four out of eleven isolates associated with the 2002 U.S. outbreak isolates had complete genomes and of those, three contained plasmids. There were no plasmids reported for the 1998 U.S. outbreak isolates and the South African outbreak isolates, so we used BLAST analysis to compare their shotgun genomes against all *L. monocytogenes* plasmids deposited in the NCBI database to identify possible plasmid contigs. We then calculated the total length of these plasmid contigs for identification.

Tip-dated phylogenetic analysis using BEAST

Among 39 genomes, the 686 SNPs were determined by the CFSAN SNP pipeline, and the SNPs were used for a tip-dated phylogeny using BEAST v2.6.2 **[68]**. To find the best supportive model, three different clock models (strict clock, relaxed clock exponential, and relaxed clock log normal) were tested with three different priors: coalescent constant population, coalescent exponential population, and coalescent Bayesian skyline. For each run, the Hasegawa–Kishino–Yano (HKY) substitution model for nucleotide evolution was selected with four gamma categories and kappa parameter of 4.0. The Markov chain Monte Carlo (MCMC) was set to 100 million and the result (tracelog and treelog) was recorded every 2000 runs. Each combination was compared based on (i) the mean marginal likelihood values and (ii) the convergence with the effective sample size (ESS) value using Tracer v1.7.1 **[69]**. Then, five replicate runs of the best clock model and prior combination were performed, and the log and tree files were combined using LogCombiner v2.6.2 **[68]** with a 10% burn-in. TreeAnnotator v2.6.2 **[68]** was used to summarize the combined tree file as a maximum clade credibility (MCC) tree with common ancestor node heights and the tree was visualized using Figtree v1.4.4 (https://beast.community/figtree).

Results

Complete genomes generated by long-read and short-read sequencing

Seventeen isolates of *L. monocytogenes* CC6 recovered from seven RTE meat and poultry processing facilities (A, B, C, D, E, F, and G) in the U.S. between 2002 and 2009 (Table 1) were sequenced by Illumina MiSeq using the whole genome shotgun strategy. Among them, five isolates, one each from five different facilities, were selected to be sequenced by ONT MinION. Upon successful sequencing of both Illumina and ONT, Unicycler used both in a hybrid assembly pipeline to generate a complete circular contig of the chromosome, with an additional circular contig if a plasmid was present (**Table II-3**). Genome coverage of the complete chromosomes was 63.9× for OB020621, 18.5× for OB030029, 91.4× for OB040119, 86.7× for OB050226, and 38.4× for OB080183; the plasmid coverage was 54.3× for OB030029, 304.8× for OB040119, 166.8× for OB050226, and 60.0× for OB080183. After each genome was polished by its corresponding shotgun data, the total number of nucleotides changed was 35 for OB020621, 29 for OB030029, 25 for OB040119, 137 for OB050226, and 36 for OB080183.

Table II-3. Five complete chromosome and plasmid genomes of *L. monocytogenes* CC6 isolates from five different facilities. The genomes were assembled using both long-reads and short-reads using Unicycler assembler.

Isolate	Facility	Circular	Number of contigs	Length of chromosome (bp)	Length of plasmid (bp)
OB020621	A	Yes	1	2,949,231	Not found
OB030029	B	Yes	2	2,994,628	55,801
OB040119	C	Yes	2	2,953,956	55,803
OB050226	D	Yes	2	2,953,451	55,798
OB080183	E	Yes	2	2,908,728	90,421

Single Nucleotide Polymorphism (SNP) Analysis by Center for Food Safety and Applied Nutrition (CFSAN) SNP Pipeline

CFSAN SNP Pipeline was used to identify SNPs among the 17 non-outbreak isolates (**Table II-1**) and 22 shotgun-sequenced isolates from three different listeriosis outbreaks that occurred in the U.S. and South Africa associated with RTE meat or poultry products (**Table II-2**). The complete genome of OB020621 (Facility A) was chosen to be the reference for the chromosomal sequences of these isolates. A phylogenetic tree showed monophyletic clades consisting of isolates from the same outbreak or processing facility (**Figure II-1a**).

The non-outbreak isolates from each cluster differed by less than 20 SNPs: the two isolates from Facility A differed by 2 SNPs; the two isolates from Facility C differed by 6 SNPs; the six isolates from Facility D showed a difference by 4 to 20 SNPs (median, 10 SNPs); and the four isolates from Facility F differed by 1 to 4 SNPs. Therefore, isolates from each facility likely belonged to the same strain. The number of SNPs between clusters representing different facilities ranged from 45 to 247 SNPs.

Nine food and clinical isolates from the 1998 listeriosis outbreak were grouped into one clade containing two subclades: one subclade containing five isolates (H7355, H7738, H7762, H7961, and H7962) differing by ≤11 SNPs, and the other subclade containing four isolates (H7550, H7557, H7596, and H7969) differing by ≤3 SNPs. The isolates from the two subclades differed by 15 to 21 SNPs. Eight isolates from the 2002 listeriosis outbreak similarly formed a single clade containing two subclades, one subclade contained six food and clinical isolates that differed by

≤5 SNPs while the second subclade contained two environmental isolates (J1815 and J1816) that differed by 3 SNPs. The two subclades had 60 to 65 SNP differences. Lastly, the five isolates associated with the South African outbreak formed a single clade with ≤3 SNPs. Among these three outbreaks, the South African outbreak isolates were relatively close to the 1998 U.S. outbreak isolates, and they differed by 66 to 80 SNPs. In contrast, the 2002 U.S. outbreak isolates and the 1998 U.S. outbreak isolates differed by 227 to 249 SNPs. Comparing the relationship among all CC6 isolates analyzed in this study, the non-outbreak isolates from Facilities A, E, F, and G, the 1998 U.S. outbreak isolates, and the South African outbreak isolates formed Clade I (**Figure II-1a**), and they differed by 43 to 83 SNPs; the isolates recovered from Facilities B, C, and D, and the 2002 U.S. outbreak isolates formed Clade II (**Figure II-1a**), and they differed by 63 to 113 SNPs. These two clades differed by 218 to 257 SNPs.

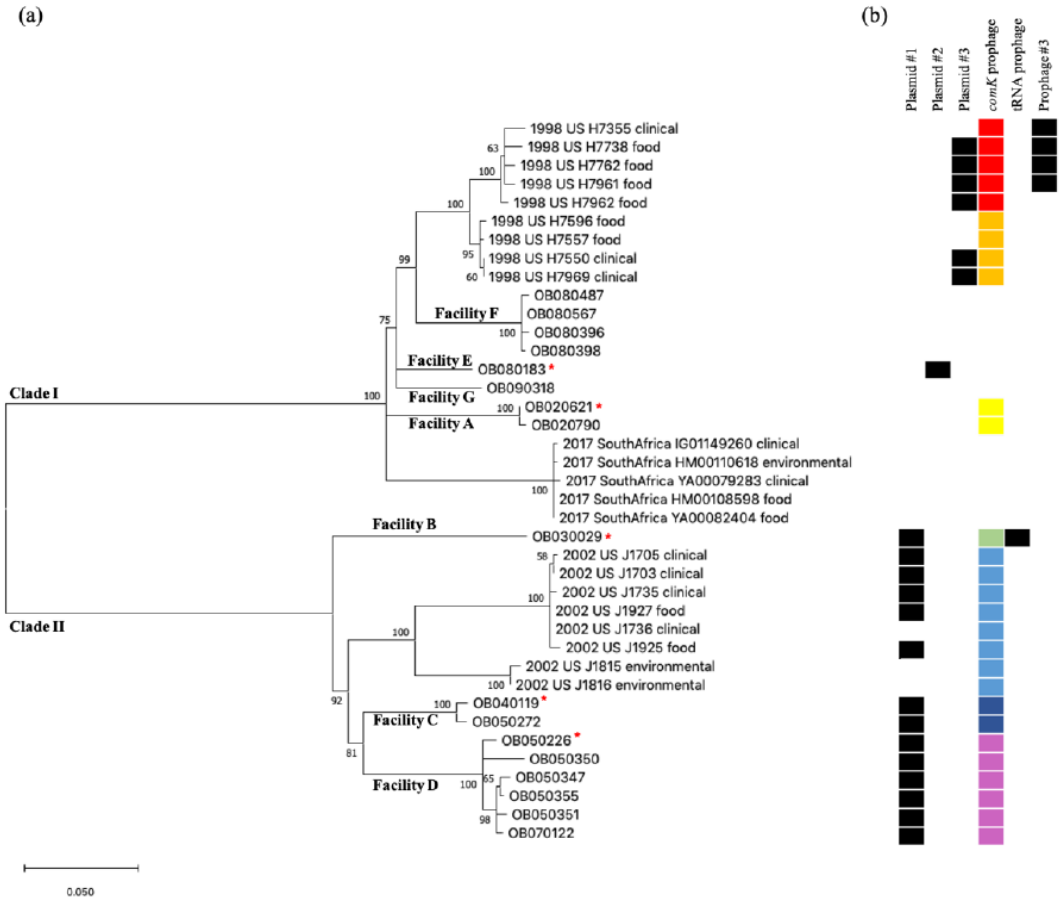


Figure II-1. Phylogenetic tree constructed by Center for Food Safety and Applied Nutrition (CFSAN) single nucleotide polymorphism (SNP) pipeline. Bootstraps values (n = 1,000) labeled on each branch (a). The isolates sequenced by Oxford Nanopore Technologies MinION are indicated (*). (b) The presence or absence of each plasmid and prophage is indicated (presence: black/absence: blank). Identical *comK* prophages are indicated with the same color.

Identification and Characterization of Prophages from Shotgun and Complete Genomes

Among isolates from the seven facilities, isolates from Facilities A, B, C, and D had "intact" prophages on the shotgun genomes predicted by PHASTER (PHAge Search Tool–Enhanced Release). OB020621 and OB020790 (Facility A) each had one 41.2 Kb "intact" prophage; OB030029 (Facility B) had two "intact" prophages, 42.5 Kb and 43.1 Kb; OB040119 and OB050272 (Facility C) each had one "intact" prophage, 95.4 Kb and 80.9 Kb, respectively; OB050226 and OB050350 (Facility D) each had one 42.6 Kb "intact" prophage while OB050347, OB050351, OB050355, and OB070122 (Facility D) each had two "intact" prophages, 27.9 Kb and 31.1 Kb. Isolates from Facilities E, F, and G had no "intact" prophages predicted, and the predicted "incomplete" prophages were all in the middle of shotgun contigs, meaning these prophages were not a fraction of a prophage that was split into multiple incompletely assembled contigs. An intact, complete *comK* gene (609 bp) was present in the isolates from Facilities E, F, and G, confirming that no *comK* prophage was present in these isolates. Each one of the "intact" prophages of the isolates from Facilities A, B, C, and D was determined as a possible *comK* prophage if it contained at least one piece of the disrupted *comK* gene (423 bp or 189 bp). These results led us to subject one isolate from each of Facilities A, B, C, and D to long-read sequencing to obtain its complete genome and to identify the complete prophage region.

Among the five complete genomes of isolates from Facilities A, B, C, D, and E, OB020621 from Facility A had one "intact" prophage of 54.3 Kb; OB030029 from Facility B had two "intact" prophages of 43.1 Kb and 53.9 Kb; OB040119 from

Facility C had one 45.8 Kb "intact" prophage; and OB050226 from Facility D had one "intact" prophage of 53.9 Kb. In contrast, OB080183 from Facility E did not have any "intact" prophage predicted. We then identified complete or disrupted *comK* genes in these genomes and found that except for one (43.1 Kb) of the two prophages in OB030029, all other "intact" prophages were *comK* prophages. The other "intact" prophage (43.1 Kb) found on the OB030029 complete genome was the only non-*comK* prophage, and we identified tRNA-Lys as the insertion site. There was no "incomplete" prophage predicted from any of the complete genomes.

We subsequently used the *comK* gene to modify the beginning and end positions of PHASTER-predicted prophage from complete genomes, and the lengths were ~41K bp for all *comK* prophages. The PHASTER-predicted prophage regions from complete genomes were always ~6 Kb to ~14 Kb longer than those prophage regions modified using the *comK* gene. PHASTER predicted two attachment sites (attL and attR) to determine the beginning and end position of a prophage, and the attachment sites were always slightly upstream or downstream of the disrupted *comK* gene. The NCBI annotation of these predicted regions showed the actual phage-like proteins were always between the disrupted *comK* gene locations.

For the four isolates from Facilities A, B, C, and D, the predicted "intact" prophage of a complete genome always corresponded to an "intact" prophage of the shotgun genome of the same isolate. However, the "intact" prophage predicted from the shotgun genome of OB040119 (Facility C, 95.4 Kb) was twice the size of its corresponding region predicted from the complete genome of OB040119 (45.8 kb), which was finally determined to be a 40.2 Kb *comK* prophage. Detailed sequence

analysis revealed that the 95.4 Kb prophage from the OB040119 shotgun genome contained two identical copies of the 40.2 Kb *comK* prophage. We subsequently used SKESA to assemble these short reads, and the same region contained only one copy of the 40.2 Kb *comK* prophage, confirming that SPAdes assembly had an error. When we used the disrupted *comK* gene to modify the start and end positions of these "intact" prophages from the SKESA-assembled shotgun genome, the final length was consistent with the *comK* prophage identified from its corresponding complete genome.

There were no "incomplete" prophages predicted from any of the five complete genomes. In contrast, a 22.8 Kb or 16.9 Kb "incomplete" prophage was predicted from each of the shotgun genomes of these five isolates and shared 100% sequence identity among them. This "incomplete" prophage was in the middle of a shotgun contig for all five isolates, and the examination of protein annotations showed that there was no integrase and the total length of phage-like proteins was only 9.4 Kb. This 22.8 Kb or 16.9 Kb region was also present in the complete genomes, but it was not predicted as an "incomplete" or "intact" prophage when the complete genomes were analyzed by PHASTER. This was consistent with our approach of not considering "incomplete" prophage regions predicted in the middle of shotgun contigs.

We then compared the prophages predicted from isolates that were subjected to long-read sequencing with the isolates from the same facility that were only subjected to short-read sequencing and identified a few differences. Specifically, for the Facility D isolates, the 27.9 Kb "intact" prophage of four isolates (OB050347,

OB050351, OB050355, and OB070122) corresponded to the ~41 Kb *comK* prophage identified in long-read and short-read sequenced OB050226. Detailed examination of sequences revealed that the 27.9 Kb "intact" prophage was at the end of a shotgun contig in all four isolates, indicating that the *comK* prophage might be split into multiple shotgun contigs in these four isolates. We subsequently identified another portion (~13 Kb) of this *comK* prophage in another contig of each isolate, and that region was predicted as an "incomplete" prophage. The other shotgun-sequenced isolate, OB050350, from Facility D had a 42.6 Kb "intact" prophage predicted on a single contig, which corresponded to the *comK* prophage of OB050226 and the actual length was also ~41Kb after we modified its start and end locations using the disrupted *comK* gene. Therefore, for Facility D, among the five isolates subjected only to shotgun-sequencing, the *comK* prophage was intact in one isolate and split into at least two contigs in the remaining four isolates. Thus, these results showed that the "intact" prophages predicted by PHASTER might not be the actual complete prophage when the predicted region was located at the end of a shotgun contig. Another difference was observed among isolates in Facility C. Similar to the shotgun genome of OB040119, the 80.9 Kb "intact" prophage in the shotgun genome of OB050272 contained two identical copies of ~41 Kb *comK* prophage identified from the complete genome of OB040119, and SKESA assembly from the short reads only showed one copy of the *comK* prophage, indicating that SPAdes shotgun genome assembly had an error.

In summary, the prophage prediction from the complete genome was slightly more accurate than those from the shotgun genomes because (i) shotgun genomes

could have assembly errors caused by repeated regions in prophages, (ii) "incomplete" prophage predicted from shotgun genomes could be due to incompletely assembled fragments of an "intact" prophage and therefore inaccurate, and (iii) a prophage could be split into multiple shotgun contigs, resulting inaccurate predictions for both parts of the prophage. A total of four *comK* prophage regions were identified in the non-outbreak isolates from four facilitates (A, B, C, and D). Interestingly, isolates recovered from the same facility had a 100% identical (0 SNPs) *comK* prophage (**Figure II-1b**). The four *comK* prophages from different facilities (A, B, C, and D) had significant sequence variations (**Figure II-2**). The *comK* prophage of isolates in Facility A was most divergent from the other *comK* prophages of isolates from other facilities (i.e., 17% to 36% of BLAST alignment coverage (AC) with above 88% of sequence identity (SI)) (**Figure II-2**). In contrast, the prophages from Facilities B, C, and D had a much higher similarity (i.e., 74% to 88% AC and 91% to 96% SI) (**Figure II-2**). We also performed gene-by-gene BLAST comparisons; the four *comK* prophages contained 63 to 65 genes, the prophage from Facility A shared 14 to 30 genes with prophages from other facilities, and the prophages from Facilities B, C, and D shared 43 to 54 genes. These suggested that possible horizontal gene transfer or prophage replacement occurred during a short-term evolution.

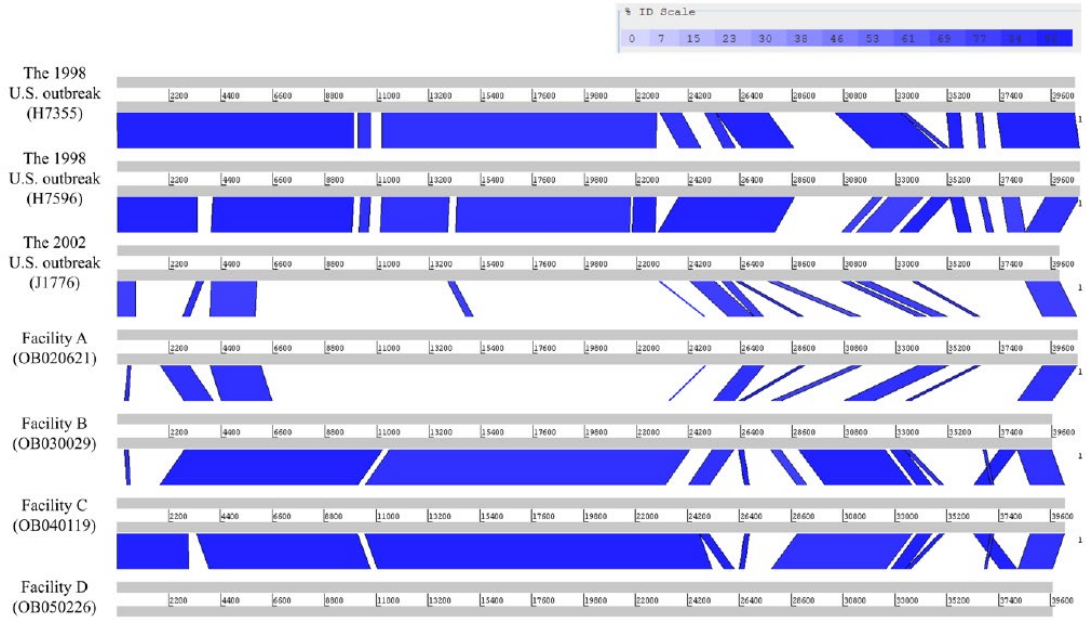


Figure II-2. Artemis Comparison Tool comparison of seven complete *comK* prophages found among the 39 *L. monocytogenes* CC6 isolates. The *L. monocytogenes* CC6 isolates recovered from Facilities A, B, C, and D shared the same *comK* prophage. Five isolates belonged to the 1998 U.S. outbreak shared an identical *comK* prophage indicated as H7355, while four isolates also belonged to the 1998 U.S. outbreak harbored the *comK* prophage indicated as H7596. All nice isolates associated with the 2002 U.S. outbreak contained an identical *comK* prophage indicated as J1776. No *comK* prophage was determined among the 2017 South African isolates.

Identification of prophages harbored in *L. monocytogenes* isolates from three outbreaks associated with RTE meat or poultry products

No complete genomes of the isolates from the 1998 U.S. outbreak were available in the NCBI database. PHASTER was used to predict prophages from genomes assembled with short-read shotgun sequencing data. The prophage profiles of the nine 1998 isolates corresponded to the two subclades of the phylogenetic tree generated by CFSAN SNP Pipeline (Figure 1). The first subclade contained five isolates, H7355, H7738, H7762, H7961, and H7962. PHASTER predicted a 43.5 Kb "intact" prophage and a 48.3 Kb "intact" prophage from both H7355 and H7738. The predicted prophages were all in the middle of relatively large contigs (i.e., 335 Kb and 620 Kb for H7355 and 195 Kb and 620 Kb for H7738, respectively), which increased our confidence in the prophage predictions from these shotgun genomes. The 43.5 Kb prophages from H7355 and H7738 were the same (100% AC and 100% SI) and harbored the disrupted *comK* gene near both ends. We subsequently modified it to be 40,606 bp using the disrupted *comK* gene. The 48.3 Kb prophages from H7355 and H7738 were also identical (100% AC and 100% SI), however, we could not determine the exact insertion site and referred to it as Prophage #3 hereinafter. PHASTER predicted a 41.1 Kb "intact" prophage at the end of a shotgun contig from each of H7762, H7961, and H7962. Only one part of the disrupted *comK* gene was found to flank this 41.1 Kb "intact" prophage region, and this 41.1 Kb "intact" prophage region was nearly identical to the 40,606 bp *comK* prophage (99% AC and 100% SI) from H7355 and H7388. Thus, we used the *comK* prophage from H7355 and H7388 as the reference and located the complete *comK* prophage from H7961,

H7762, or H7962 that was split into two contigs. We subsequently determined that the *comK* prophages from all five isolates were identical (**Figure II-1b**). This is another example showing that the "intact" prophages predicted from shotgun genomes may not represent the complete prophage. Furthermore, PHASTER predicted prophage #3 in addition to the *comK* prophage from H7961 and H7762. H7962 did not have prophage #3, corresponding to the subclades of the phylogenetic tree (**Figure II-1**).

The second subclade contained four isolates, H7596, H7550, H7557, and H7969. PHASTER predicted a 41.3 Kb "intact" prophage in the middle of a relatively long contig (i.e., 335 Kb) from H7596. We then identified the *comK* gene as the insertion site and used the disrupted *comK* gene to modify the prophage region to be 40,815 bp. PHASTER predicted a 29.7 Kb "intact" prophage at the end of a contig from H7550, H7557, and H7969 with identical sequences. This 29.7 Kb prophage was flanked by one part of the disrupted *comK* gene and partially aligned with the 40,815 bp *comK* prophage (72% AC and 100% SI). We then used the 40,815 bp *comK* prophage as the reference and identified the other part (~10 Kb) of this *comK* prophage from H7550, H7557, and H7969, which were predicted as "incomplete" prophages by PHASTER. The ~10 Kb was flanked by the second part of the disrupted *comK* gene and was always in a different contig from that containing the 29.7 Kb region. We subsequently determined that the *comK* prophages from each isolate within this subclade were identical (**Figure II-1b**).

The two *comK* prophages (40,606 bp and 40,815 bp) from the two subclades of the 1998 U.S. outbreak isolates had 81% AC to each other with 99% SI and they

shared 51 genes out of the total of 65 genes (**Figure II-2**), even though they were associated with the same outbreak. These suggested that possible horizontal gene transfer or prophage replacement occurred during a short-term evolution.

Four complete genomes from the 2002 U.S. outbreak isolates were available to increase the confidence of obtaining complete prophages with PHASTER analysis. All four complete genomes had one "intact" prophage predicted with the *comK* gene as the insertion site. We modified this prophage region to ~40 Kb using the disrupted *comK* gene. The eight shotgun genomes also had one "intact" prophage with the *comK* gene as the insertion site and there were no "incomplete" or "intact" prophages at the ends of a shotgun contig. The *comK* prophages predicted were identical (100% AC and 99% SI) in all complete and shotgun genomes (**Figure II-1b**).

PHASTER predicted one "intact" prophage of 62.2Kbfromthe one complete genome (HM00113468) available for the South African outbreak. Two "intact" prophages of 62.2 Kb and 31 Kb were predicted in the middle of relatively large shotgun contigs (141 Kb and 148 Kb) from YA00079283 and we likely identified the complete region of each prophage. The 62.2 Kb "intact" prophages predicted from the complete genome and the shotgun genome corresponded to each other (98% AC and 97% SI) and this prophage was split into multiple contigs in the other four shotgun genomes. We also identified a ~28.5 Kb "intact" prophage region and a ~18.8 Kb "incomplete" prophage regions that were at the end of shotgun contigs in all shotgun-sequenced isolates, and thus they could be only portions of prophage(s), and we could not identify the entire prophage(s). The 31 Kb, ~28.5 Kb, and ~18.8 Kb prophages were only predicted from shotgun genomes, and they were not identified from the

complete genome of HM00113468. Thus, gain or loss of prophages likely occurred in South African outbreak isolates. The ~18.8 Kb prophage region appeared to have tRNA-Lys as the insertion site, while the other three prophages (62.2 Kb, 31 Kb, and ~28.5 Kb) were not inserted into *comK* or near tRNA-Lys. All six South African isolates had an intact complete *comK* gene (609 bp) in their genomes. Hence, there was no *comK* prophage integrated into the South African isolates.

Identification of plasmids on both non-outbreak and outbreak isolates

Out of the four plasmids identified by ONT MinION sequencing, the plasmids (hereinafter referred to as Plasmid #1) (~56 Kb) from OB030029, OB040119, and OB050226 (Facility B, C, and D, respectively) were 99% identical to each other by BLAST alignment. Comparison of Plasmid #1 with *L. monocytogenes* plasmids in the NCBI database showed that it was nearly identical to the plasmid found in J1776 isolate of the 2002 U.S. outbreak (100% AC and 99% SI). The 90,421 bp plasmid (hereinafter referred to as Plasmid #2) found in the complete genome of OB080183 (i.e., the only isolate in Facility E) was different from Plasmid #1 (17% AC and 99% SI). Comparison with Plasmid #2 sequences deposited in NCBI showed that it had 98% AC and 99% SI to the plasmid of strain LM-F-75 (NCBI Accession number: KY613765, 91,243 bp).

BLAST of Plasmid #1 against 17 non-outbreak shotgun genomes from all facilities showed that this plasmid was present in the isolates belonging to Facilities B, C, and D. This plasmid was split into 9 to 11 contigs in each shotgun genome. BLAST of Plasmid #2 against 17 shotgun genomes showed that it was only present in OB080183, split into five different contigs of its shotgun genome. Prediction by

PlasmidFinder-2.0 and BLAST comparison against *L. monocytogenes* plasmids deposited in the NCBI database did not identify additional plasmids in the 17 non-outbreak isolates.

Among the four complete genomes of the 2002 U.S. outbreak (J1776, J1816, J1817, and J1926) deposited in NCBI, three except for J1816 had the same plasmid (~56 Kb, 100% AC and 99% SI among different isolates) which was nearly identical to Plasmid #1 (**Figure II-3a**). Isolate J1816 did not have any plasmids. Additionally, BLAST of Plasmid #1 against shotgun-sequenced isolates associated with this outbreak showed that five shotgun genomes (J1703, J1705, J1735, J1925, and J1927) harbored this plasmid. The other two shotgun genomes (J1736 and J1815) did not have this plasmid. This was also consistent with PlasmidFinder-2.0 predictions and BLAST analysis compared with published *L. monocytogenes* plasmids, which determined that J1736 and J1815 did not have any plasmids.

No complete genomes belonging to the 1998 U.S. listeriosis outbreak were deposited in the NCBI database. Among the nine shotgun genomes of isolates associated with the 1998 U.S. outbreak, six isolates were determined to harbor a plasmid by PlasmidFinder-2.0 and BLAST analysis. BLAST analysis of these shotgun genomes with Plasmid #1, Plasmid #2, and plasmids published in NCBI showed that all isolates harbored the same 82 Kb plasmid (hereinafter referred to as Plasmid #3), which aligned with the Plasmid #2 (~91 Kb) with 89% AC and 99% of SI (**Figure II-3b**). Plasmid #3 also aligned with the plasmid of strain LM-F-131 (NCBI Accession number: NZ_CM009923.1, 81,666 bp) with ~92% AC and 99% SI. A benzalkonium chloride (BC) tolerance gene cassette, *bcrABC,* was found in both

Plasmid #2 and #3. There were no plasmids reported or found in the five shotgun genomes of the South African outbreak by PlasmidFinder-2.0. BLAST searches of the shotgun genomes against published *L. monocytogenes* plasmids genomes in the NCBI database also could not find any plasmids within the genomes of South African outbreak isolates.

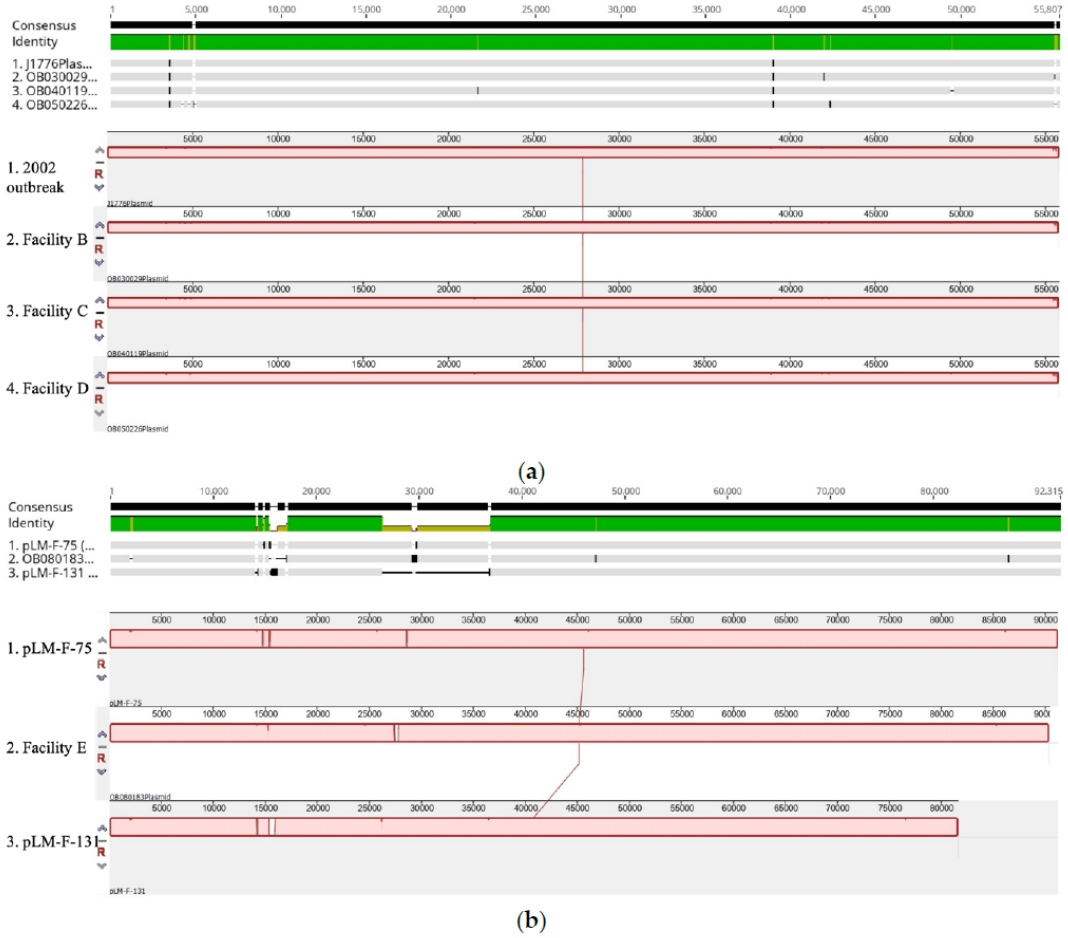


Figure II-3. Mauve alignment of plasmids. (a) Plasmid #1: this plasmid (GenBank ID: CP006612.1) was found in J1776 of the 2002 U.S. outbreak isolates. The complete plasmid genome of Facilities B, C, and D isolates showed 100% alignment coverage (AC) and 99% sequence identity (SI) with pJ1776. (b) Plasmid #2 and #3: plasmid of LM-F-75 (GenBank ID: KY613765.1) from the NCBI database showed 98% AC and 99% SI to Plasmid #2 from OB080183 (Facility E). The strain LM-F-131 (GenBank ID: QADR01000015.1) has a plasmid with ~92% AC and 99% SI to Plasmid #3 found in the six shotgun genomes of the 1998 U.S. outbreak. The plasmid of LM-F-131 showed 100% of SI but 89% AC to Plasmid #2 (~90 Kb).

Tip-Dated Phylogenetic Analysis on Both Non-Outbreak and Outbreak Isolates Using Bayesian Evolutionary Analysis by Sampling Trees (BEAST)

Based on the marginal likelihood value and the effective sample size (ESS) values, the best fitting model for the 39 *L. monocytogenes* CC6 isolates was the strict clock model with coalescent constant population prior. The mean substitution rate was 3.1×10^7 (95% highest posterior density, HPD, 1.6×10^7 to 4.6×10^7) nucleotide substitutions per site per year. The maximum clade credibility (MCC) tree (**Figure II-4**) was generated from five independent runs, and the time of most recent common ancestor (tMRCA) of various CC6 isolates was estimated: (i) July 1968 (95% HPD, January 1961 to June 1974) for isolates from the 1998 U.S. outbreak, (ii) February 1950 (95% HPD, January 1928 to November 1967) for isolates from the 2002 listeriosis outbreak, (iii) May 2015 (95% HPD, December 2012 to April 2017) for five South African outbreak isolates, and (iv) March 1816 (95% HPD, January 1716 to February 1896) for the entire 39 *L. monocytogenes* CC6 isolates.

According to the MCC tree, the 39 isolates formed two major clades, Clade I and Clade II (**Figure II-4**). The tMRCA for the Clade I containing the 1998 U.S. outbreak isolates, South African outbreak isolates, and eight non-outbreak isolates (Facilities A, E, F, and G) were estimated in March 1950 (95% HPD, March 1924 to May 1970). The other, Clade II was comprised of the 2002 U.S. outbreak isolates and nine non-outbreak isolates (Facilities B, C, and D) and the estimated tMRCA was in November 1925 (95% HPD, January 1889 to April 1955).

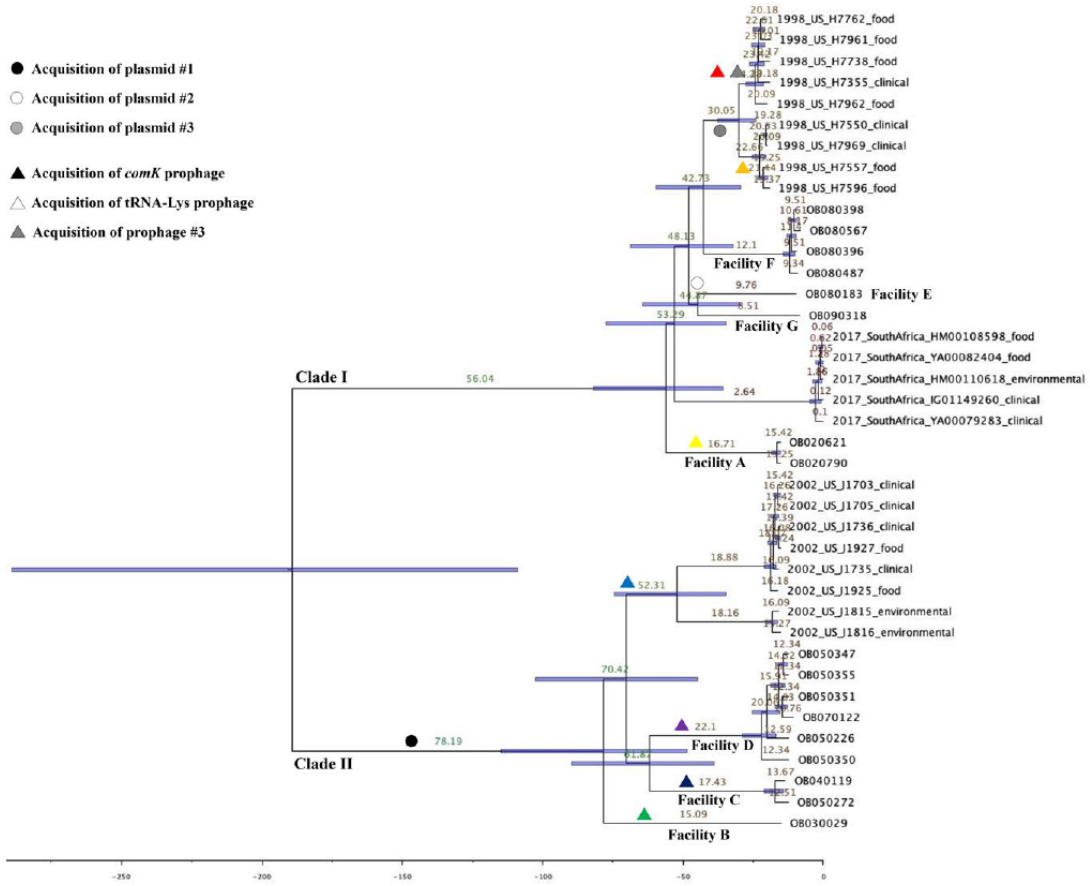


Figure II-4. The maximum clade credibility (MCC) tree of 39 CC6 isolates from Bayesian evolutionary analysis by sampling trees (BEAST) analysis. The mean heights were labeled on each branch and its 95% highest posterior density (HPD) was displayed in a horizontal bar. The possible acquisition events of plasmids and prophages are displayed; the color code of *comK* prophage matches Figure II-1. The time of the most recent common ancestor (tMRCA) was estimated: (a) July 1968 (95% HPD, January 1961 to June 1974) for the 1998 U.S. outbreak isolates, (b) February 1950 (95% HPD, January 1928 to November 1967) for the 2002 U.S. outbreak isolates, (c) May 2015 (95% HPD, December 2012 to April 2017) for the South African outbreak isolates, and (d) March 1816 (95% HPD, January 1716 to February 1896) for the entire 39 CC6 isolates.

Discussion

Previous genetic studies have speculated that MGEs acquired by horizontal gene transfer may enhance the resistance of *L. monocytogenes* to disinfection and increase the risk of persistence or repeated contamination in food processing facilities **[26, 37, 48, 70]**. These studies were based on shotgun and PCR amplicon sequencing data to determine specific genes or portions of MGEs because complete MGE sequences are difficult to identify by shotgun sequencing due to their large size and the presence of repetitive sequence elements. In recent years, long-read sequencing platforms have emerged and enabled us to obtain complete genomes **[56]**. Using a combination of both short-read and long-read sequencing platforms allowed us to generate accurate, complete genomes with speed.

We employed ONT MinION and Illumina MiSeq sequencing platforms to analyze 17 *L. monocytogenes* CC6 isolates recovered from RTE meat or poultry processing facilities, not known to be associated with listeriosis outbreaks. All isolates were subjected to Illumina MiSeq sequencing and, based on MGEs analysis of shotgun data, five isolates were further subjected to ONT MinION sequencing. In addition, 22 shotgun sequencing data and five complete genomes of the isolates belonging to three CC6 listeriosis outbreaks associated with RTE meat or poultry products were included to investigate the genetic relationship of *L. monocytogenes* CC6 strains that may have persisted in RTE meat or poultry processing facilities. The acquisition of MGEs via horizontal gene transfer could introduce highly variable sites with new genetic resistance features to *L. monocytogenes* genomes during survival and growth in the food matrix or in food processing environment **[71]**. Thus, MGE

profiling could be relevant to differentiate closely related strains and also useful to trace the evolutionary relationship of *L. monocytogenes* strains, especially within the same ECs, CCs, or strains that do not have many variations on their backbone genomes **[71]**. In our study, we focused on plasmids and chromosome-borne prophages to determine distinct genetic patterns of *L. monocytogenes* CC6 isolates associated with listeriosis outbreaks and isolates that might have been persistent or repeatedly contaminated RTE meat or poultry processing facilities in the U.S. We compared shotgun contigs to all plasmids published in the NCBI database and did not find any plasmids that might be integrated into chromosomes. The total length of plasmid contigs was the same as the length of the extrachromosomal plasmid identified by long-read sequencing. For the 1998 U.S. outbreak isolates, we do not have complete genomes, but the extrachromosomal plasmid of one of the outbreak isolates was isolated and characterized previously based on PCR amplicon sequencing and shotgun sequencing **[42]**.

The complete sequences of *comK* prophage determined using the complete genomes could serve as references to help identify the presence or absence of each complete prophage region in shotgun genomes. Prophage predictions directly from shotgun genomes were mostly consistent with prophage predictions from complete genomes. Specifically, the *comK* prophages predicted from the shotgun genomes were mostly present in one contig and corresponded to the disrupted *comK* gene that was in the same contig. We were able to modify the start and end positions of all PHASTER-predicted *comK* prophage using the disrupted *comK* gene to obtain the exact prophage region encoding prophage-like proteins. However, in the case of the

1998 U.S. outbreak, no complete genomes were available, but in several isolates, the contigs containing the *comK* prophage were very long (e.g., 195 Kb, 335 Kb, or 620 Kb), which increased our confidence of correctly identifying the complete, intact *comK* prophage. Another inconsistent PHASTER prediction from the complete genomes and from the shotgun genomes was found on a few of Facility D isolates and the 1998 U.S. outbreak isolates. For four of the isolates associated with Facility D, the *comK* prophage was split into multiple contigs with two contigs containing 80% of the entire *comK* prophage and each contig containing one part of the disrupted *comK* gene (423 bp or 189 bp); one part of the *comK* prophage was predicted as "intact" and another part was predicted as "incomplete" by PHASTER. The remaining 20% of the *comK* prophage was split into at least three other contigs. For three of the isolates associated with the 1998 U.S. outbreak, the *comK* prophage was split into only two contigs, with each contig containing one part of the disrupted *comK* gene; one part of the *comK* prophage was predicted as "intact" and the other part was predicted as "incomplete" by PHASTER. Lastly, the two non-outbreak isolates from Facility C showed inconsistent PHASTER prediction between shotgun and complete genomes due to the assembler error. Each shotgun genome of two non-outbreak isolates from Facility C assembled by SPAdes created two copies of the same prophage, whereas SKESA allowed more accurate prophage predictions by PHASTER. This highlighted the value of using long-read sequencing since the completeness and accuracy of the assembly had a remarkable effect on the correct identification of chromosome-borne prophages in bacterial genomes.

Chapter III: Prevalence and genetic diversity of *L. monocytogenes* recovered from the surface of whole fresh avocados collected in 2014

Abstract

Whole fresh fruits have become an emerging concern for *L. monocytogenes* contamination. In 2014, a survey was conducted by the U.S. Food and Drug Administration (FDA) to determine the prevalence of *L. monocytogenes* on the exterior and interior of whole fresh avocados. In this study, we analyzed whole-genome sequencing data of 135 *L. monocytogenes* strains recovered from the avocado skin samples during the 2014 survey using core genome multi-locus sequencing typing (cgMLST) and single-nucleotide polymorphism (SNP) analyses. *In silico* MLST analysis classified the isolates into lineages I (n=53, 39%) and II (n=82, 61%) and further grouped them into 22 clonal complexes (CCs) and three singletons. The most prevalent CCs were CC14 of lineage II (n=38, 28%), CC392 of lineage I (n=16, 12%) and CC412 of lineage II (n=12, 9%). The common serogroups were IIa (n=82, 61%), IIb (n=50, 37%) and IVb (n=3, 2%). The presence of major genes and genomic islands associated with *L. monocytogenes* virulence was identified among the isolates. *Listeria* pathogenic islands (LIPI) associated with hypervirulence were present in 22% for LIPI-3 and 9% for LIPI-4 of the 135 strains. Stress survival islet 1 (SSI-1) was present in 32% of the 135 strains, and SSI-2 was not found in any isolate. Full

length of *inlA* encoding internalin A was found in all 135 strains. The presence of major genes contributing to the survival and persistence of *L. monocytogenes* was determined. Comparative analysis of the *L. monocytogenes* isolates recovered from avocado skin samples highlights that whole fresh avocados can be served as a food vehicles for several CCs of *L. monocytogenes,* including hypervirulent clones.

Introduction

L. monocytogenes is widely distributed in natural environmental sources such as soil and water, leading to frequent recovery in fruits and vegetables [8]. Due to its ability to survive under refrigeration temperatures and form biofilms on dry surfaces, *L. monocytogenes* has been isolated from the surfaces of whole intact produce [72]. Since 2010, intact fruit commodities such as cantaloupes, stone fruits, and caramel apples have emerged as new food sources for *L. monocytogenes* and caused multistate listeriosis outbreaks. In 2011, a listeriosis outbreak associated with consuming contaminated cantaloupes affected 28 states with 147 illnesses and 33 deaths [73]. In 2014, a multistate listeriosis outbreak and recalls occurred and were associated with contaminated stone fruits such as whole peaches, nectarines, plums, and pluots [74]. Another multistate listeriosis outbreak associated with contaminated caramel apples occurred, affecting 12 states and causing 35 illnesses and seven deaths [75]. The emerging incidence of *L. monocytogenes* contamination on tree fruits, possibly occurring during harvest and transportation, introduced the importance of understanding the prevalence and genetic diversity of *L. monocytogenes* in such conditions to have a better strategy to track the transmission and contamination sources and ultimately reduce the potential risk of *L. monocytogenes* contamination risk.

Although no listeriosis outbreaks were linked to *L. monocytogenes* on the surface of whole fresh avocados, the commodity could be a food vehicle for *L. monocytogenes* transmitted from environments to consumers. In 2014, a survey was carried out by the FDA on whole fresh avocados to determine the prevalence of *L.*

monocytogenes on the exterior and interior of the commodities **[76]**. The FDA surveillance assignment typically designates a sample as a batch of 10 subsamples (i.e., ten individual avocados in this assignment) from the same lot. A total of 1,615 whole fresh avocado samples (i.e., 16,150 avocados) were collected and analyzed during the survey. A soak method was used to analyze *L. monocytogenes* and from the skins of 361 avocado samples (i.e., 3,610 avocados), and these samples were thereafter referred to as skin samples. The pulp of 1,254 samples (i.e., 12,540 avocados) was aseptically taken out for detection of *L. monocytogenes,* and these samples were thereafter referred to as pulp samples. Each sample was tested using a wet composite scheme, which first generated two composites per sample by pooling primary enrichment cultures of 5 subsamples into a composite. They were then pooled into the same secondary enrichment broth, followed by Vitek Immunodiagnostic Assay System (VIDAS) rapidity screening and cultural confirmation. When VIDAS generated a positive result for a composite, the five primary enrichment cultures corresponding to the presumptive positive composite were then individually analyzed. This survey was the first nationwide survey of *L. monocytogenes* on whole fresh avocados. Overall, *L. monocytogenes* was detected in 64 out of 361 avocado skin samples (18%), and three out of 1,254 avocado pulp samples (0.24%) **[76]**. During this avocado survey, whole-genome sequencing (WGS) was performed on all *L. monocytogenes* isolates recovered, and the sequencing data were deposited into the GenomeTrakr database.

In this chapter, WGS data of *L. monocytogenes* isolated from the surface of avocados were analyzed to 1) assess the clonal diversity of *L. monocytogenes*, 2)

identify major genetic elements associated with virulence potential and stress resistance and 3) determine strain relatedness for identifying the possible transmission events of *L. monocytogenes.*

Materials and Methods

Data collection of *L. monocytogenes* isolates

Among the 1,615 avocado samples (i.e., 16,150 avocados) analyzed during the 2014 survey, 361 avocado samples (i.e., 3,610 avocados) were subjected to a soak method to determine the prevalence of *L. monocytogenes* on the surface of avocado fruits. Each avocado sample comprising ten avocados was pooled to generate two composites. If a composite (i.e., five avocado subsamples) tested positive for *L. monocytogenes*, the original enrichment culture of each subsample was further tested. Thus, each sample could yield up to 12 isolates, two from composites and ten from individual subsamples. During this survey, a total of 64 avocado samples yielded *L. monocytogenes* isolates, representing an overall prevalence of 18% among the 361 avocado skin samples [76]. As a part of the surveillance, whole-genome sequencing was performed on *L. monocytogenes* isolates and the sequencing data was deposited in the GenomeTrakr database. In this chapter, a total of 250 *L. monocytogenes* isolates from 84 composites and 166 subsamples were retrieved, and the paired-end reads were *de novo* assembled using Qiagen CLC Genomics Workbench 21.0.5 (Aarhus, Denmark). The quality of *L. monocytogenes* genomes was assessed using QUAST v5.0.2 [77] based on a total length of 3.0 ± 0.3 Mb, length of N50 and a number of contigs less than 300.

Core-genome MLST and determination of duplicated isolates

A molecular typing based on a 1,827-cgMLST scheme [28] built-in Ridom© SeqSphere+ software (Ridom© GmbH, Münster, Germany) was performed on the 250

L. monocytogenes genomes. Each genome was evaluated if it contained >95% of the "good" cgMLST targets, which represents that the alignment against a reference target had at least 80% identity, gene length with ± three triplets, and no ambiguities or frameshifts [78]. A minimum spanning tree (MST) was generated using the pairwise allelic differences (AD) based on 1,827 cgMLST targets.

Due to the composite scheme, the 250 isolates may include duplicates from the same sample. Different isolates from the same sample likely belonged to the same strain, and thus we considered them duplicates and only chose one isolate to represent that strain for the sample. Clustering of the 250 isolates on MST was used to depict duplicates using two thresholds (i.e., 7 and 12) of maximum AD. The threshold of 7 AD was previously proposed to identify genetically related isolates [79], but due to the small dataset size, a threshold of 12 was included to detect putative duplicated isolates [46]. When the neighboring isolates on MST showed ≤ 7 AD and belonged to the same sample, they were considered duplicates; thus, we chose one as a representing isolate for the sample for further analyses. When the neighboring isolates had between 7 and 12 maximum AD and belonged to the same sample, we considered these as putative duplicated isolates; these putative duplicated isolates were further analyzed using the CFSAN SNP pipeline [59] to generate SNP distance. When the SNP distance among the putative duplicated isolates was ≤ 20 SNPs, the isolates were considered as the same strain [80, 81]; thus, one representing isolate was selected from each sample for further analyses. Lastly, when the isolates showed >12 AD, they were considered different strains even if recovered from the same sample. Consequently, 135 *L. monocytogenes* strains were selected in this study for further

phylogenetic and statistical analyses. A phylogenetic tree was constructed by a single-linkage algorithm based on the pairwise AD generated by the 1,827-cgMLST scheme in Ridom SeqSphere^{+}.

Determination of PCR-serogroups, clonal complexes (CCs), sequence types (STs) and lineages.

In silico PCR-serogroups identification was performed using Ridom SeqSphere^{+} to identify the presence of four PCR genetic markers described by Doumith et al. [20]. The profiles were used to assign one of four major *L. monocytogenes* serogroups: IIa (1/2a and 3a), IIb (1/2b, 3b and 7), IIc (1/2c and 3c) and IVb (4b, 4d and 4e). STs were determined using *in silico* MLST scheme based on seven housekeeping genes of *L. monocytogenes* using Ridom SeqSphere^{+}. CCs were further assigned based on the definition given by Ragon et al. [82]. The MLST profiles were curated according to the Institut Pasteur MLST *Listeria* database (http://bigsdb.pasteur.fr/listeria/listeria.html). The lineages of the 135 strains were determined based on the PCR serogroups and the phylogeny clades.

The presence or absence of major virulence genes, premature stop codons (PMSCs) in the inlA gene, genes involved in stress resistance and plasmids.

Genetic elements previously described as associated with virulence or stress resistance phenotypes of *L. monocytogenes* [45, 46, 83, 84] were assessed in the 135 genomes using the nucleotide BLAST. The nucleotide BLAST alignment with ≥70% sequence coverage and ≥80% sequence identity was applied as parameters. The presence of PMSCs in the *inlA* gene was determined by extracting the *inlA* sequences

from each genome using CLC Genomic Workbench. The *inlA* sequences were aligned using MEGA X, including the *inlA* gene (lmo0433) from *L. monocytogenes* EGD-e as a reference. The presence of plasmid sequences in shotgun genomes was determined using two approaches. First, the *repA* gene that encodes plasmid replication initiation protein RepA in *Listeria* [85] was determined in each genome using the nucleotide BLAST using the same parameters listed above. Secondly, each contig of the shotgun genomes was aligned against the complete sequences of 68 *Listeria* plasmids deposited in the GenBank as of April 2021. When a contig showed ≥60% sequence coverage and ≥70% sequence identity, we considered the contig belonged to a plasmid.

Results

Description of *L. monocytogenes* isolates.

Among the 64 positive avocado skin samples containing 128 composites (2 composites per avocado sample), 84 composites yielded *L. monocytogenes* isolates. Subsequent analysis of enrichment cultures of individual subsamples yielded 166 *L. monocytogenes* isolates among 420 subsamples (5 subsamples per composite). The 64 positive avocado skin samples were initially grown in three regions, 28 from California (44%), three from Chile (5%) and 33 from Mexico (52%). The collection of the 64 avocado samples occurred in multiple geographical regions: 44 positive avocado samples, including domestically grown (i.e., California) and imported avocados (i.e., Chile or Mexico), were locally collected in 12 U.S. states, whereas the remaining 20 positive samples grown in Chile or Mexico were collected at ports of the U.S. entry [76].

Although an avocado sample consisted of 10 subsamples, a positive sample may yield up to 12 isolates from both subsamples and composites (10 from subsamples and two from composites); for instance, samples 871510 and 802499 each yielded 11 isolates, consisting of two composite isolates and nine subsample isolates. The pairwise analysis based on the 1,827-cgMLST scheme determined that 112 isolates had ≤ 7 AD (i.e., indicating isolates likely belonged to the same strain) from at least one of the other isolates from the same avocado sample. Including all these isolates in the final analysis would bring bias to the calculation of prevalence of each genotype of *L. monocytogenes*; thus, we considered these 112 isolates as

duplicates and only retained one for subsequent analyses. Three additional isolates were determined as duplicates when we employed the maximum 12 AD threshold and 20 SNPs differences based on the whole genomes. After removing the 115 duplicated isolates, the 135 strains represented the genetic diversity of *L. monocytogenes* associated with avocado skin samples (**Table III-1**). In summary, among 64 avocado skin samples, 29 avocado samples (45%) had one strain per sample, 16 avocado samples (25%) had two strains per sample, ten avocado samples (16%) had three strains per sample, and nine avocado samples had more than three strains per sample (**Table III-1**). Two samples had notably high level of diversity, sample 871510 grown in Mexico contained eight strains belonging to five CCs while sample 802499 grown in California contained six strains belonging to five CCs (**Table III-1**).

PCR-serogroups and phylogenetic tree of the 135 *L. monocytogenes* strains.

Three *Listeria* serogroups, IIa (1/2a and 3a), IIb (1/2b, 3b and 7) and IVb (4b, 4d and 4e), were identified among the 135 strains. The prevalence of serogroups were serogroup IIa (n=82, 61%), serogroup IIb (n=50, 37%) and serogroup IVb (n=3, 2%) (**Table III-2**). Among the 66 isolates recovered from domestically grown avocados, 46 isolates (70%) belonged to serogroup IIb. Among the 69 isolates recovered from imported avocados, 63 isolates (91%) belonged to serogroup IIa (**Table III-2**). According to the phylogenetic tree constructed using the pairwise AD of cgMLST (**Figure III-1**), lineage I (n=53, 39%) and lineage II (n=82, 61%) were identified based on the serogroup identification and the phylogenetic clades. The pairwise AD among lineage I isolates (n=53) ranged from 5 to 1,301 (median: 1,177), whereas

lineage II isolates ranged from 0 to 1,493 (median: 1,350). The pairwise AD between lineages I and II isolates ranged from 1,768 to 1,796.

Table III-1. Summary of the 135 *L. monocytogenes* strains selected in this study. The sample ID was designated by the 64 positive avocado sample names followed by composite (C) or subsample (S) ID.

Sample ID [a]		FDA accession	Lineage	Serogroup	ST [b]	CC [c]	Geographical location [d]	
							Growing	Sampling
412696	S5	CFSAN022369	II	IIa	365	CC14	Mexico	TX
412700	S4	CFSAN022781	I	IIb	392	CC392	Mexico	TX
412700	S9	CFSAN022782	II	IIa	365	CC14	Mexico	TX
433341	C1	CFSAN022783	II	IIa	365	CC14	Mexico	CA
433341	S4	CFSAN022786	II	IIa	7	CC7	Mexico	CA
433341	S5	CFSAN022787	II	IIa	365	CC14	Mexico	CA
433341	S8	CFSAN022788	II	IIa	573	CC573	Mexico	CA
562911	S1	NA[e]	I	IIb	2546	CC288	CA	CA
562911	S3	NA	I	IIb	288	CC288	CA	CA
562911	S5	NA	II	IIa	412	CC412	CA	CA
773872	S6	CFSAN025756	I	IIb	224	CC224	CA	CO
773872	S4	CFSAN025754	II	IIa	412	CC412	CA	CO
802499	S2	CFSAN022391	I	IIb	288	CC288	CA	CA
802499	S3	CFSAN022392	I	IIb	392	CC392	CA	CA
802499	S4	CFSAN022393	I	IIb	59	CC59	CA	CA
802499	S8	CFSAN022396	I	IIb	288	CC288	CA	CA
802499	S1	CFSAN022389	II	IIa	918	CC918	CA	CA
802499	S10	CFSAN022390	II	IIa	412	CC412	CA	CA
802506	S9	CFSAN061848	I	IIb	288	CC288	CA	CA
802506	S1	CFSAN025760	II	IIa	412	CC412	CA	CA
802506	S5	CFSAN025761	II	IIa	412	CC412	CA	CA
802506	S6	CFSAN061849	II	IIa	412	CC412	CA	CA
802792	S1	CFSAN025763	II	IIa	412	CC412	CA	CA

831281	S2	CFSAN022702	II	IIa	206	CC14	CA	VA
831844	S2	CFSAN018334	II	IIa	1044	ST1044	CA	WA
841045	S1	CFSAN022713	I	IIb	224	CC224	CA	AZ
841045	S3	CFSAN064043	I	IIb	392	CC392	CA	AZ
841045	S7	CFSAN022714	I	IIb	363	CC5	CA	AZ
841045	S8	CFSAN022712	I	IIb	59	CC59	CA	AZ
841045	S10	CFSAN022716	I	IIb	59	CC59	CA	AZ
846140	S10	CFSAN061839	I	IIb	392	CC392	CA	CO
847171	C2	CFSAN018332	II	IIa	14	CC14	Chile	NJ
847171	S7	CFSAN018333	II	IIa	7	CC7	Chile	NJ
849740	S7	CFSAN023060	II	IIa	412	CC412	CA	IL
849740	S10	CFSAN023062	II	IIa	412	CC412	CA	IL
851203	S1	CFSAN064049	I	IIb	363	CC5	CA	CA
851203	S6	CFSAN064050	I	IIb	392	CC392	CA	CA
854428	C2	CFSAN023064	I	IIb	224	CC224	CA	UT
854428	S4	CFSAN033508	I	IIb	59	CC59	CA	UT
854428	S9	CFSAN023068	I	IIb	363	CC5	CA	UT
854428	S3	CFSAN033507	II	IIa	918	CC918	CA	UT
854691	S6	CFSAN025765	I	IIb	392	CC392	CA	MI
856009	C1	CFSAN059051	I	IVb	4	CC4	CA	TX
856009	S3	NA	I	IIb	392	CC392	CA	TX
856009	S4	NA	I	IIb	224	CC224	CA	TX
856011	S10	CFSAN023072	II	IIa	365	CC14	Mexico	TX
857042	S3	NA	I	IIb	288	CC288	CA	CA
857042	S5	NA	I	IIb	224	CC224	CA	CA
859897	S1	CFSAN022813	II	IIa	918	CC918	CA	WA
860005	S5	CFSAN025031	I	IIb	5	CC5	Mexico	CA
860005	S2	CFSAN064051	II	IIa	365	CC14	Mexico	CA

860005	S10	CFSAN025033	II	IIa	365	CC14	Mexico	CA
861279	S3	CFSAN022374	I	IIb	782	CC2	CA	TX
861559	S5	CFSAN023074	II	IIa	412	CC412	CA	CO
862597	S6	CFSAN018245	II	IIa	37	CC37	Chile	Chile
863017	S5	CFSAN018248	II	IIa	398	CC19	Chile	Chile
863232	S6	CFSAN025018	II	IIa	7	CC7	Mexico	Mexico
863232	S8	CFSAN025019	II	IIa	365	CC14	Mexico	Mexico
863766	S5	CFSAN022817	I	IIb	288	CC288	CA	AZ
863766	S6	CFSAN022818	I	IIb	224	CC224	CA	AZ
863766	S7	CFSAN022819	I	IIb	288	CC288	CA	AZ
864162	S10	CFSAN022399	II	IIa	365	CC14	Mexico	Mexico
865491	S1	CFSAN022791	II	IIa	1108	CC321	Mexico	TX
865491	S4	CFSAN022792	II	IIa	365	CC14	Mexico	TX
865491	S5	CFSAN022793	II	IIa	365	CC14	Mexico	TX
866014	S5	CFSAN025768	I	IIb	392	CC392	Mexico	MA
866014	S7	CFSAN025769	II	IIa	365	CC14	Mexico	MA
866014	S10	CFSAN025770	II	IIa	365	CC14	Mexico	MA
866278	S6	CFSAN023078	I	IIb	59	CC59	CA	CA
866278	S7	CFSAN023079	I	IIb	363	CC5	CA	CA
866278	S9	CFSAN023081	I	IIb	59	CC59	CA	CA
866278	S10	CFSAN023082	I	IIb	363	CC5	CA	CA
866278	S3	CFSAN023077	II	IIa	91	CC14	CA	CA
866425	S8	CFSAN033521	I	IIb	363	CC5	CA	TX
868395	C2	CFSAN033513	I	IIb	736	CC736	CA	UT
868395	S3	CFSAN033514	I	IIb	392	CC392	CA	UT
868395	S7	CFSAN033515	I	IIb	392	CC392	CA	UT
868396	S4	CFSAN033517	I	IIb	392	CC392	CA	UT
868396	S8	CFSAN024973	I	IIb	392	CC392	CA	UT

868396	S9	CFSAN024974	I	IIb	224	CC224	CA	UT
868396	S7	CFSAN024972	II	IIa	412	CC412	CA	UT
868553	S9	CFSAN023901	II	IIa	1118	ST1118	Mexico	Mexico
868576	S4	CFSAN023903	II	IIa	365	CC14	Mexico	Mexico
868831	S4	CFSAN023913	II	IIa	7	CC7	Mexico	Mexico
868831	S6	CFSAN023917	II	IIa	1321	CC1056	Mexico	Mexico
868831	S8	CFSAN023921	II	IIa	365	CC14	Mexico	Mexico
868939	S8	CFSAN023929	II	IIa	365	CC14	Mexico	Mexico
869095	S10	CFSAN025778	II	IIa	365	CC14	Mexico	MA
869096	S5	CFSAN025780	II	IIa	364	CC364	Mexico	MA
869097	S1	CFSAN025782	II	IIa	1056	CC1056	Mexico	MA
869098	S1	CFSAN025784	II	IIa	365	CC14	Mexico	MA
869296	S5	CFSAN025785	I	IIb	392	CC392	CA	CA
869296	S1	CFSAN061847	II	IIa	918	CC918	CA	CA
869839	S5	CFSAN061837	II	IIa	412	CC412	CA	OK
870166	S3	CFSAN024981	II	IIa	570	CC570	Mexico	Mexico
870170	S2	CFSAN024998	II	IIa	1126	ST1126	Mexico	Mexico
870170	S10	CFSAN024999	II	IIa	364	CC364	Mexico	Mexico
870204	S2	CFSAN024986	II	IIa	365	CC14	Mexico	Mexico
870204	S10	CFSAN024988	II	IIa	573	CC573	Mexico	Mexico
870353	S6	CFSAN024990	II	IIa	365	CC14	Mexico	Mexico
870353	S7	CFSAN025000	II	IIa	573	CC573	Mexico	Mexico
870405	S2	CFSAN024991	II	IIa	946	CC945	Mexico	Mexico
870405	S4	CFSAN025003	II	IIa	365	CC14	Mexico	Mexico
870420	S6	CFSAN025005	II	IIa	7	CC7	Mexico	Mexico
870420	S7	CFSAN064078	II	IIa	365	CC14	Mexico	Mexico
870830	S2	CFSAN025009	II	IIa	573	CC573	Mexico	Mexico
870830	S3	CFSAN025010	II	IIa	37	CC37	Mexico	Mexico

870830	S5	CFSAN025011	II	IIa	365	CC14	Mexico	Mexico
870857	S7	CFSAN025013	II	IIa	7	CC7	Mexico	Mexico
870860	S3	CFSAN025016	II	IIa	365	CC14	Mexico	Mexico
870860	S5	CFSAN025017	II	IIa	1301	CC14	Mexico	Mexico
871048	S10	CFSAN024977	II	IIa	365	CC14	Mexico	TX
871250	S4	CFSAN061844	I	IIb	363	CC5	CA	CA
871319	S7	CFSAN025791	I	IVb	1292	CC6	Mexico	CA
871319	S2	CFSAN025790	II	IIa	365	CC14	Mexico	CA
871510	S1	CFSAN025794	II	IIa	365	CC14	Mexico	TX
871510	S3	CFSAN025795	II	IIa	365	CC14	Mexico	TX
871510	S4	CFSAN025796	II	IIa	365	CC14	Mexico	TX
871510	S6	CFSAN025797	II	IIa	1108	CC321	Mexico	TX
871510	S8	CFSAN025799	II	IIa	365	CC14	Mexico	TX
871510	S9	CFSAN025800	II	IIa	37	CC37	Mexico	TX
871510	C1	CFSAN025792	II	IIa	573	CC573	Mexico	TX
871510	S7	CFSAN025798	II	IIa	946	CC945	Mexico	TX
871808	S3	CFSAN025802	I	IVb	639	CC639	Mexico	TX
871808	S4	CFSAN025803	II	IIa	1056	CC1056	Mexico	TX
871808	S5	CFSAN025804	II	IIa	365	CC14	Mexico	TX
872232	S8	NA	I	IIb	363	CC5	CA	CA
872500	S2	CFSAN025022	I	IIb	392	CC392	Mexico	Mexico
872500	S5	CFSAN025023	II	IIa	365	CC14	Mexico	Mexico
872732	S5	CFSAN025028	II	IIa	365	CC14	Mexico	Mexico
872841	S2	NA	I	IIb	392	CC392	CA	IL
872841	S4	NA	I	IIb	392	CC392	CA	IL
872841	S8	NA	I	IIb	224	CC224	CA	IL
872841	S10	NA	I	IIb	288	CC288	CA	IL
872942	S8	CFSAN061850	II	IIa	365	CC14	Mexico	Mexico

[a] Composite (C) and subsample (S).

[b] Sequence type (ST).

[c] Clonal complex (CC).

[d] Arizona (AZ), California (CA), Colorado (CO), Illinois (IL), Massachusetts (MA), Michigan (MI), New Jersey (NJ), Oklahoma (OK), Texas (TX), Utah (UT), Virginia (VA), Washington (WA).

[e] NA; not available.

Table III-2. Serogroups of the 135 *L. monocytogenes* strains selected in this chapter.

	Number of isolates			
Growing region	**IIa (serotype 1/2a and 3a)**	**IIb (serotype 1/2b, 3b and 7)**	**IVb (serotype 4b, 4d and 4e)**	**Total**
California	19	46	1	66
Mexico	59	4	2	65
Chile	4	0	0	4
Total	82	50	3	135

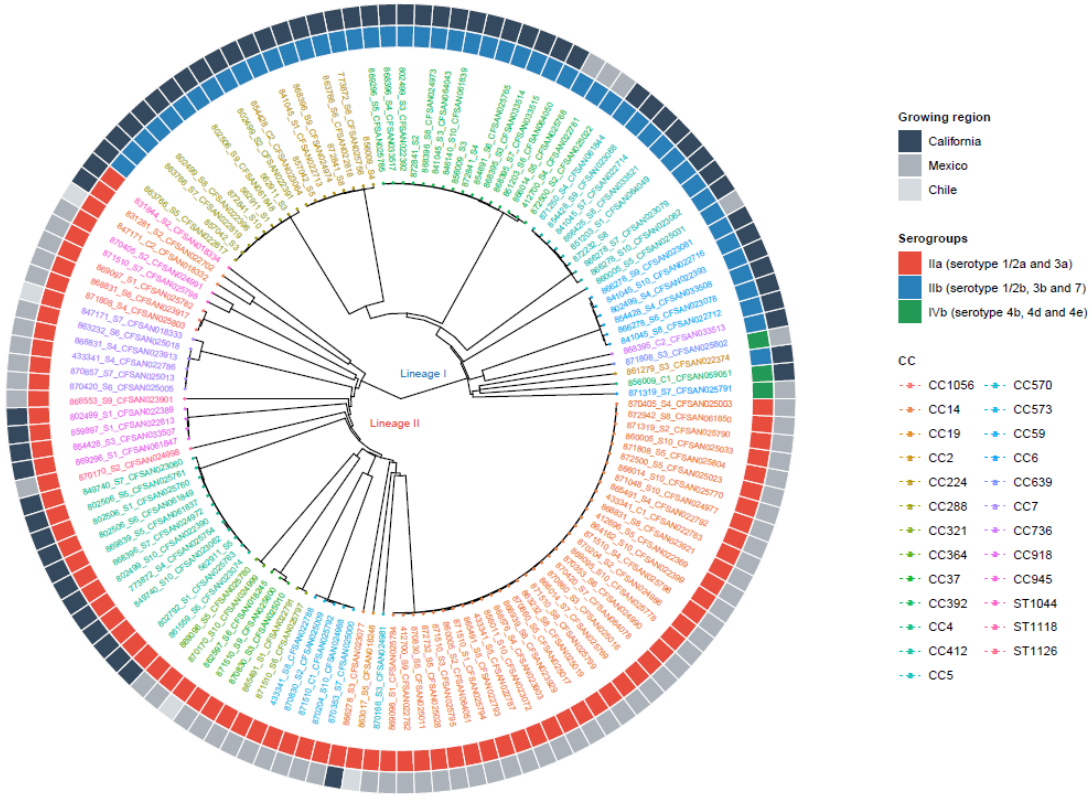


Figure III-1. Phylogeny of the 135 *L. monocytogenes* strains. The phylogenetic tree was built using a single-linkage algorithm based on the pairwise allelic differences generated by 1,827-cgMLST scheme. The inner circle represents serogroups for each isolate, and the outer circle represents the growing region for each avocado sample. The clonal complexes (CCs) for each isolate are color coded.

Description of the selected *L. monocytogenes* strains.

Overall, 22 CCs and three singletons, including 32 STs, were identified among the 135 *L. monocytogenes* strains from the avocado skin samples (**Table III-3**). Furthermore, 10 CCs were determined among the 53 lineage I strains, and 12 CCs and three singletons were determined among the 82 lineage II strains (**Table III-3 and Figure III-2A**). The most prevalent CC among the lineage I isolate was CC392 (n=16, 30%), followed by CC288 (n=9, 17%) and CC5 (n=9, 17%) (**Figure III-2**). Significantly, hypervirulent MLST clones, previously described as more frequently associated with listeriosis illness [9], were identified (i.e., CC2, CC4, and CC6) among the 135 strains. The most prevalent CC among the lineage II strains was CC14 (n=38, 46%), followed by CC412 (n=12, 15%) and CC7 (n=6, 7%) (**Figure III-2A**).

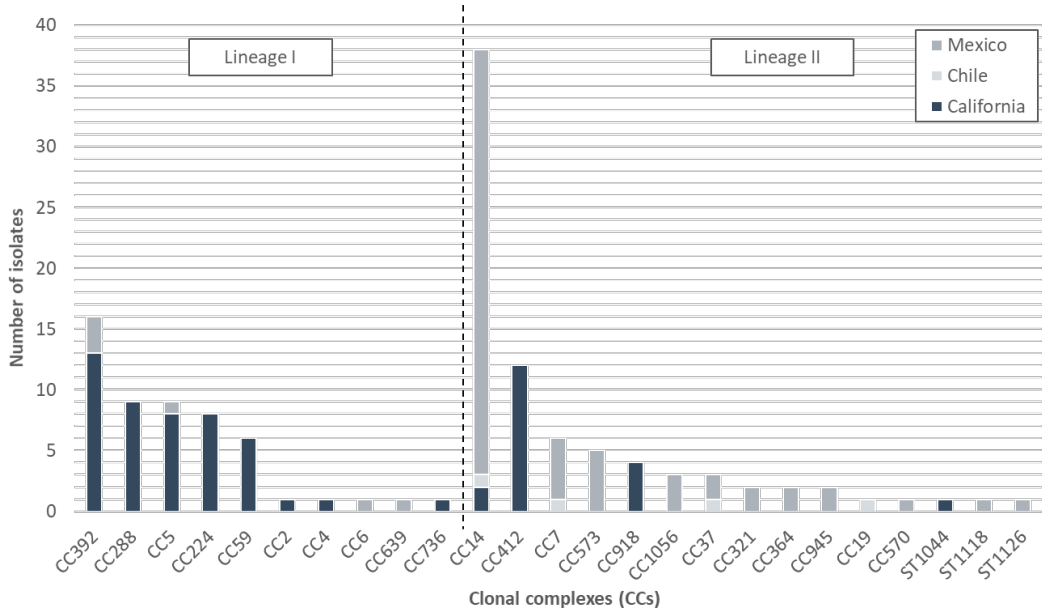
Figure III-2B shows the clonal diversity of the 135 strains per growing region: 66 strains (49%) belonged to the 28 domestically grown avocado samples (i.e., California), 65 strains (48%) belonged to the 33 avocado samples grown in Mexico and four strains (3%) belonging to the three avocado samples grown in Chile. A total of 11 CCs and one singleton were determined among the 28 domestically grown avocado samples, whereas a total of 13 CCs and two singletons were determined among the 33 avocado samples grown in Mexico (**Figure III-2B**). The CC present in the highest number of samples for each growing region was CC392 (n=13) for domestically grown avocado samples and CC14 (n=35) for avocado samples grown in Mexico.

Table III-3. The maximum allelic differences based on the 1,827-cgMLST scheme within each clonal complex (CC) and/or sequence type (ST).

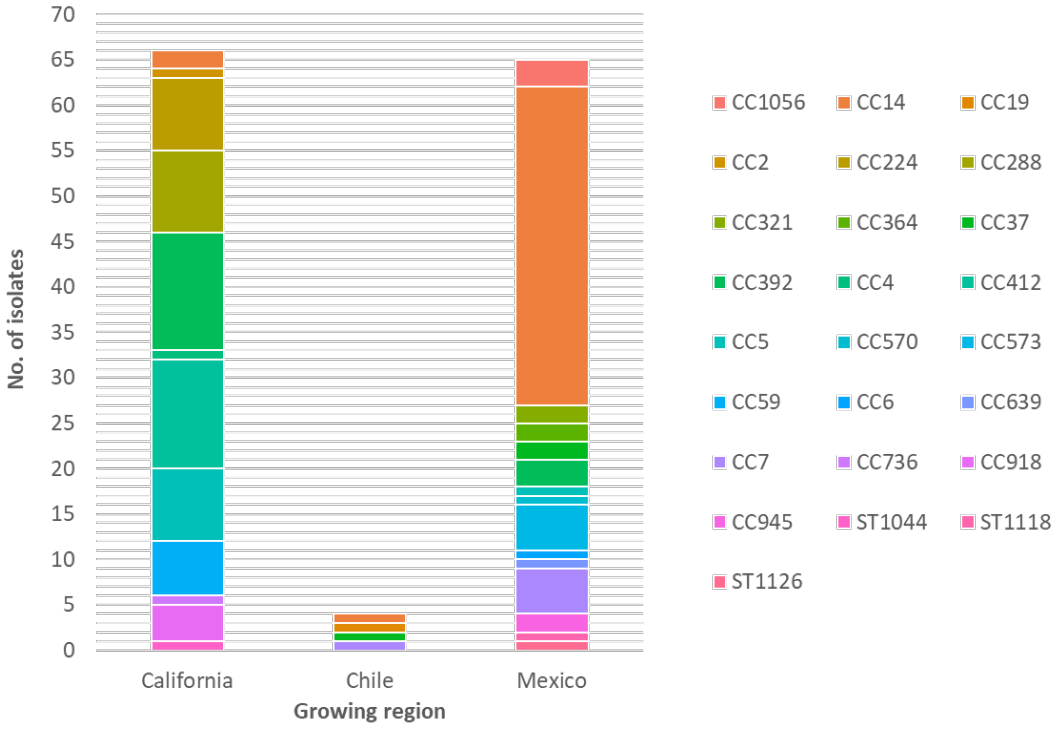
CCs	STs	No. of isolates (%)	No. of avocado samples	Maximum allelic differences (CC)	Maximum allelic differences (ST)
Lineage I					
CC392	ST392	16 (11.9%)	13	49	49
CC288	ST288	8 (5.9%)			40
	ST2546	1 (0.7%)	6	40	-[a]
CC5	ST363	8 (5.9%)	7		54
	ST5	1 (0.7%)	1	1182	-
CC224	ST224	8 (5.9%)	8	41	41
CC59	ST59	6 (4.4%)	4	51	51
CC2	ST782	1 (0.7%)	1	-	-
CC4	ST4	1 (0.7%)	1	-	-
CC6	ST1292	1 (0.7%)	1	-	-
CC639	ST639	1 (0.7%)	1	-	-
CC736	ST736	1 (0.7%)	1	-	-
Lineage II					
CC14	ST365	34 (25.2%)	27		
	ST1301	1 (0.7%)	1		31
	ST14	1 (0.7%)	1		-
	ST206	1 (0.7%)	1		-
	ST91	1 (0.7%)	1	1438	-
CC412	ST412	12 (8.9%)	9	42	42
CC7	ST7	6 (4.4%)	6	93	93
CC573	ST573	5 (3.7%)	5	29	29
CC918	ST918	4 (3.0%)	4	25	25
CC1056	ST1056	2 (1.5%)	2		29
	ST1321	1 (0.7%)	1	29	-
CC37	ST37	3 (2.2%)	3	38	38
CC321	ST1108	2 (1.5%)	2	12	12
CC364	ST364	2 (1.5%)	2	11	11
CC945	ST946	2 (1.5%)	2	195	195
CC19	ST398	1 (0.7%)	1	-	-
CC570	ST570	1 (0.7%)	1	-	-
ST1044	ST1044	1 (0.7%)	1	-	-
ST1118	ST1118	1 (0.7%)	1	-	-
ST1126	ST1126	1 (0.7%)	1	-	-

[a] No value.

(A)


Lineage I
Lineage II
Mexico
Chile
California
Number of isolates
0
5
10
15
20
25
30
35
40
CC392
CC288
CC5
CC224
CC59
CC2
CC4
CC6
CC639
CC736
CC14
CC412
CC7
CC573
CC918
CC1056
CC37
CC321
CC364
CC945
CC19
CC570
ST1044
ST1118
ST1126
Clonal complexes (CCs)

(B)


No. of isolates
0
5
10
15
20
25
30
35
40
45
50
55
60
65
70
California
Chile
Mexico
Growing region
CC1056
CC14
CC19
CC2
CC224
CC288
CC321
CC364
CC37
CC392
CC4
CC412
CC5
CC570
CC573
CC59
CC6
CC639
CC7
CC736
CC918
CC945
ST1044
ST1118
ST1126

(C)

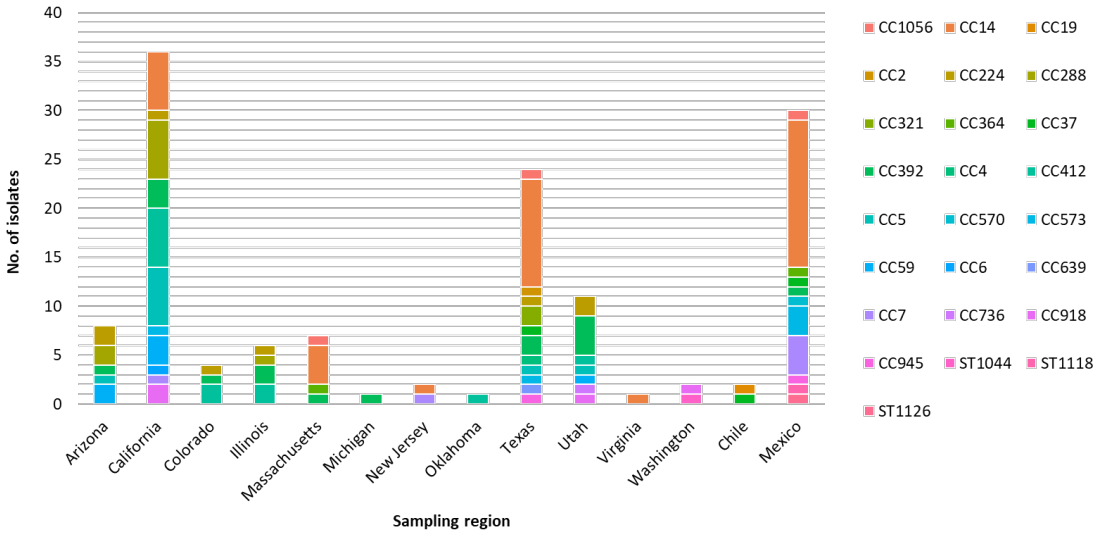


Figure III-2. Distribution of clonal complexes (CCs) and singletons among the 135 strains. A total of 22 CCs and three singletons, containing 32 STs, were identified among the 135 strains. (A) The 53 lineage I isolates were classified into 10 CCs containing 12 STs and the 82 lineage II isolates were classified into 12 CCs and three singletons containing 20 STs. The growing region of the avocado sample is indicated in different colors (dark gray: California; gray: Mexico; light gray: Chile). (B) Distribution of 22 CCs and three singletons per growing region. The Shannon's diversity index for clones found in avocado samples grown in California (n=28) was 2.16, and in Mexico (n=33) was 1.83. (C) Distribution of 22 CCs and three singletons per sampling region.

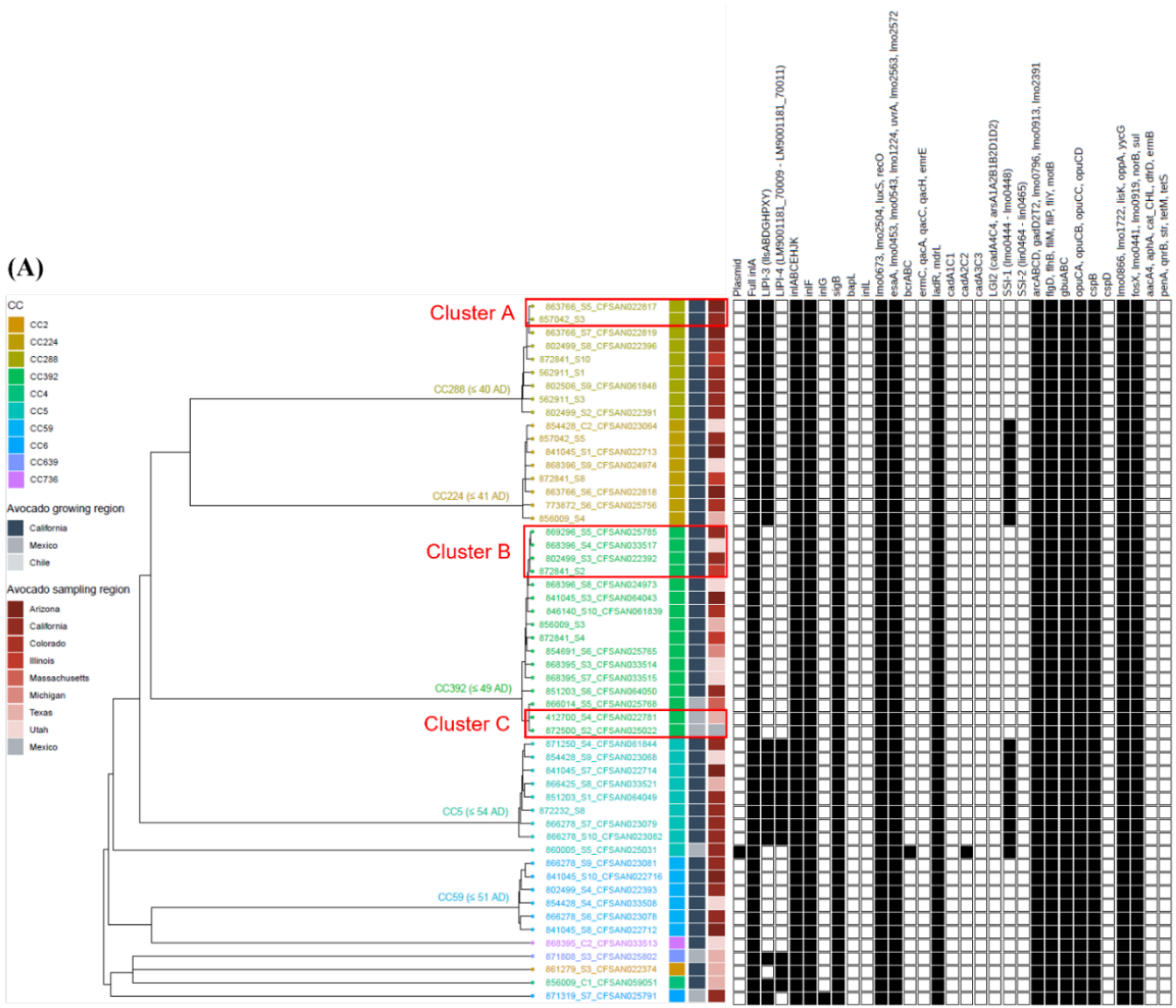
(**Figure III-2B**). The Shannon's diversity index for clonal diversity of avocado samples grown in California and Mexico was 2.16 and 1.83, respectively.

Lastly, Figure 2C shows the clonal diversity per sampling region. Among the 44 avocado samples collected in 12 local states, 103 *L. monocytogenes* strains belonging to 20 CCs and one singleton were determined and widely distributed regardless of the geographic sampling regions. Most of the sampling of 44 positive avocado samples occurred in California (n=13, 30%), consequently yielding 36 *L. monocytogenes* strains belonging to 11 CCs (**Figure III-2C**).

Phylogenetic analysis of the 135 *L. monocytogenes* strains.

L. monocytogenes strains belonging to the same CC mainly formed distinct clades, referred to as monophyletic clades hereafter, on the cgMLST-based phylogenetic tree (**Figure III-1**). Specifically, except CC5 and CC14, clones comprising more than two strains grouped into monophyletic clades with a pairwise AD ranging from 11 to 195 based on the cgMLST clustering (**Figure III-3**). The monophyletic clades contained multiple avocado samples sampled from various regions but mainly grown in the same region (**Figure III-3**). For instance, clades of CC224 (n=8, ≤ 41 AD), CC288 (n=9, ≤ 40 AD), CC5 (n=8, ≤ 54 AD), CC59 (n=6, ≤ 51 AD), CC412 (n=12, ≤ 42 AD) and CC918 (n=4, ≤ 25 AD) contained multiple avocado samples (8, 6, 7, 4, 9 and 4, respectively) grown in California (**Figure III-3**). Clades of CC573 (n=5, ≤ 29 AD), CC321 (n=2, 12 AD), CC364 (n=2, 11 AD), CC1056 (n=3, ≤ 29 AD) and CC945 (n=2, 195 AD) consisted of multiple avocado

samples (5, 2, 2, 3 and 2, respectively) grown in Mexico (**Figure III-3**). Interestingly, when the monophyletic


(A)
CC
CC2
CC224
CC288
CC392
CC4
CC5
CC59
CC6
CC639
CC736
Avocado growing region
California
Mexico
Chile
Avocado sampling region
Arizona
California
Colorado
Illinois
Massachusetts
Michigan
Texas
Utah
Mexico
Cluster A
Cluster B
Cluster C
CC288 (≤ 40 AD)
CC224 (≤ 41 AD)
CC392 (≤ 49 AD)
CC5 (≤ 54 AD)
CC59 (≤ 51 AD)
Plasmid
Full inlA
LIPI-3 (llsABDGHPXY)
LIPI-4 (LM9001181_70009 - LM9001181_70011)
inlABCEHJK
inlF
inlG
sigB
bapL
inlL
lmo0673, lmo2504, luxS, recO
esaA, lmo0453, lmo0543, lmo1224, uvrA, lmo2563, lmo2572
bcrABC
ermC, qacA, qacC, qacH, emrE
ladR, mdrL
cadA1C1
cadA2C2
cadA3C3
LGI2 (catA4C4, arsA1A2B1B2D1D2)
SSI-1 (lmo0444 - lmo0448)
SSI-2 (lin0464 - lin0465)
arcABCD, gadD2T2, lmo0796, lmo0913, lmo2391
flgD, flhB, fliM, fliP, fliY, motB
gbuABC
opuCA, opuCB, opuCC, opuCD
cspB
cspD
lmo0866, lmo1722, lisK, oppA, yycG
fosX, lmo0441, lmo0919, notB, sul
aacA4, aphA, cat_CHL, dfrD, ermB
penA, qnrB, str, tetM, tetS

SS

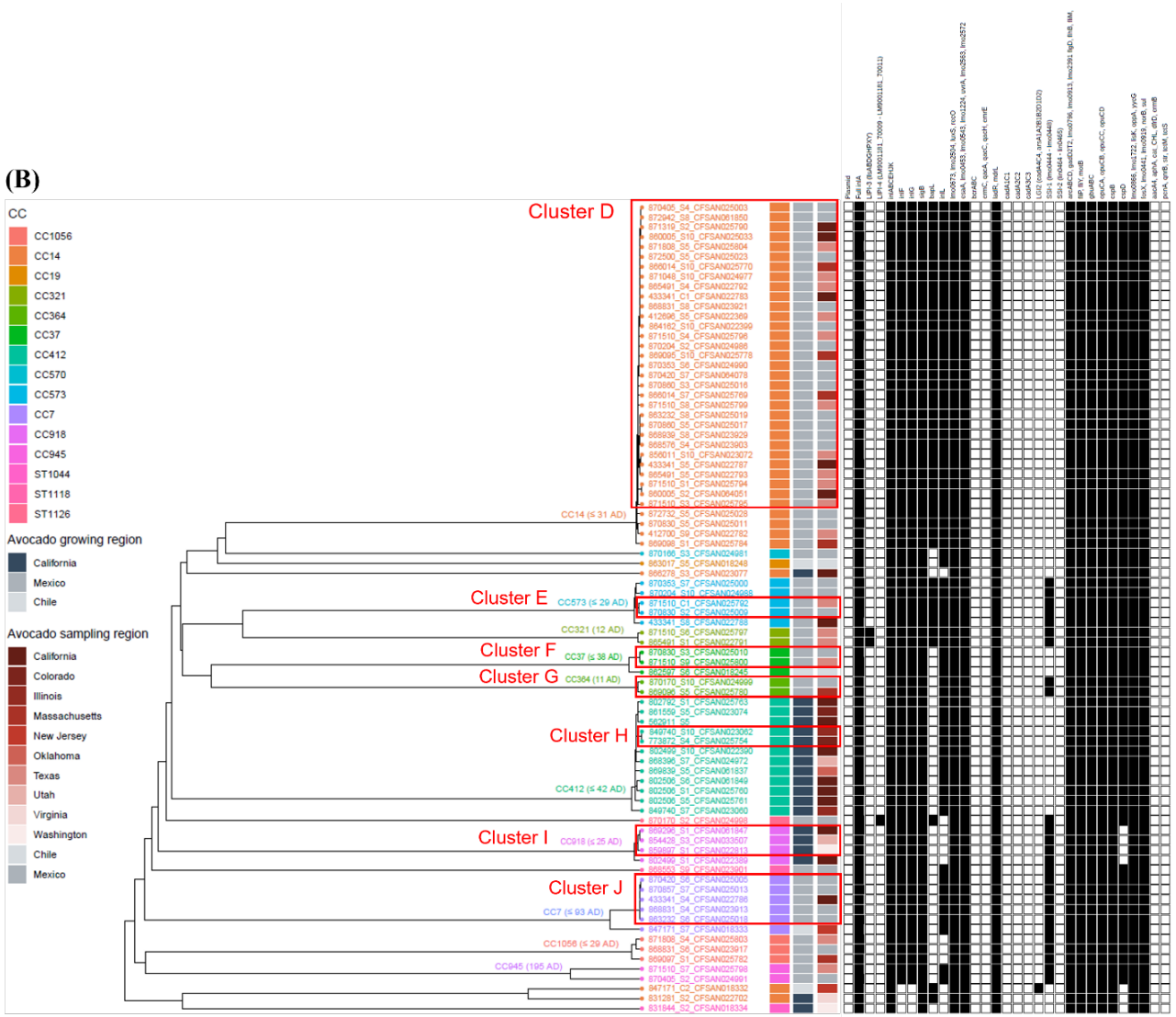

(B)
CC
CC1056
CC14
CC19
CC321
CC364
CC37
CC412
CC570
CC573
CC7
CC918
CC945
ST1044
ST1118
ST1126
Avocado growing region
California
Mexico
Chile
Avocado sampling region
California
Colorado
Illinois
Massachusetts
New Jersey
Oklahoma
Texas
Utah
Virginia
Washington
Chile
Mexico
Cluster D
Cluster E
Cluster F
Cluster G
Cluster H
Cluster I
Cluster J
CC14 (≤ 31 AD)
CC573 (≤ 29 AD)
CC321 (12 AD)
CC37 (≤ 38 AD)
CC364 (11 AD)
CC412 (≤ 42 AD)
CC918 (≤ 25 AD)
CC7 (≤ 93 AD)
CC1056 (≤ 29 AD)
CC945 (195 AD)

Figure III-3. The presence or absence of major genes associated with *L. monocytogenes* virulence or stress resistance with a phylogenetic tree. Phylogenetic trees of (A) 53 lineage I isolates and (B) 82 lineage II isolates were constructed using the pairwise allelic differences generated by 1,827-cgMLST scheme. Clusters contained isolates within a maximum allelic difference of 12 are indicated in red boxes (Clusters A – J). The presence (black) or absence (white) of the selected genes, plasmid and full length of *inlA* gene is shown next to the corresponding isolate.

clades contained avocado samples grown in different regions, the strains from the same region were related (i.e., ≤ 12 AD) and formed clusters. For instance, the monophyletic clade of CC392 (n=16, ≤ 49 AD) contained 13 avocado samples grown in California (n=10) and Mexico (n=3); three strains from avocado samples grown in Mexico formed a cluster even though they were collected in different geographical regions (i.e., Texas, Massachusetts and Mexico) (**Figure III-3A**). Similarly, another monophyletic clade of CC37 (n=3, ≤ 38 AD) contained three avocado samples grown in Mexico (n=2) and Chile (n=1); two strains from avocado samples grown in Mexico formed a cluster but each was collected from Mexico and Texas (**Figure III-3B**). Lastly, a monophyletic clade of CC7 (n=6, ≤ 93 AD) also contained six avocado samples grown in Mexico (n=5) and Chile (n=1); five strains from avocado samples grown in Mexico formed a cluster when the collection occurred in Mexico and California (**Figure III-3B**).

Two clones, CC5 and CC14, appeared in more than one monophyletic clade, referred to as paraphyletic clades hereafter, on the cgMLST-based phylogenetic tree (**Figure III-1**). Among the nine CC5 strains, eight were recovered from seven domestically grown avocado samples, whereas one strain was from an imported avocado sample from Mexico (**Figure III-3A**). The strains from domestically grown avocado samples formed one clade with a maximum AD ≤ 54 (**Figure III-3A**). The one strain from an imported avocado sample was genetically distinct from other CC5 strains (1,176 ≤ AD ≤ 1,182) even though they all belonged to the same CC. Among 38 CC14 strains, three strains from two domestically grown avocado samples and one imported from Chile (i.e., CFSAN023077, CFSAN018332 and CFSAN022702) were

genetically distinct from other CC14 strains (n=35, ≤ 31 AD) from 27 avocado samples grown in Mexico (>1,243 AD, >1,412 AD and >1,432 AD, respectively). Although CFSAN023077 and CFSAN022702 were from avocado samples grown in California, they were genetically distinct (i.e., 1,458 AD) and classified into different ST (i.e., ST91 and ST206, respectively), implying different sources of contamination. Overall, there appeared to be no association between the sampling regions of the 64 positive avocado samples and the phylogenetic clusters (**Figure III-3**).

Clusters of *L. monocytogenes* strains from multiple avocado samples on the phylogenetic tree.

Clusters containing strains recovered from different avocado samples that likely were the same strain (≤ 7 AD or ≤ 12 AD and ≤ 20 SNPs) were identified among the 135 strains based on the phylogenetic analysis (**Figure III-3**). Among the 53 lineage I strains, three clusters (Clusters A–C) were identified (**Figure III-3A**): Cluster A contained two CC288 strains that differed by nine alleles and 19 SNPs and Cluster C contained two CC392 strains that differed by ten alleles and 16 SNPs. Cluster B contained four CC392 strains that differed by ≤ 10 alleles but further determined that they likely belonged to two strains based on SNPs analysis. Two strains from avocado samples 868396 and 869296 differed by 15 SNPs and two strains from avocado samples 802499 and 872841 differed by 18 SNPs. However, these two strains differed by >20 SNPs, which did not fulfill our threshold to define the same strain. Among the 82 lineage II strains, seven clusters (Clusters D – J) were identified (**Figure III-3B**): Cluster E contained two CC573 strains that differed by nine alleles and 13 SNPs, Cluster F contained two CC37 strains differed by five

alleles, Cluster G contained two CC364 strains differed by 11 alleles and 19 SNPs, Cluster H contained two CC412 strains differed by one allele, Cluster I contained three CC918 strains differed by ≤ 12 alleles and ≤ 20 SNPs, and Cluster J contained five CC7 strains differed by ≤ 12 alleles and ≤ 20 SNPs. The largest cluster identified among the 135 strains was Cluster D contained 31 CC14 strains that differed by ≤ 12 alleles between any two strains. SNPs analysis determined that these 31 strains likely belonged to three strains that differed by ≤ 20 SNPs (**Figure III-4**). Thus, 13 strains were identified from multiple avocado samples among the 135 strains.

Genetic traits associated with virulence and pathogenesis among the 135 *L. monocytogenes* strains.

Major genes associated with *L. monocytogenes* virulence and pathogenic traits were identified to reveal the virulence potential of the 135 *L. monocytogenes* strains recovered from the avocado skin samples. All strains contained LIPI-1, including *prfA, plcA, hly, mpl, actA* and *plcB*. LIPI-3 (*llsAGHXBYDP*) was found in 22% of the 135 strains (30/135), 28 strains belonging to lineage I (53%) and two strains belonging to lineage II (2%) (**Figure III-3**). LIPI-4 (*LM9005581_70009* to *LM9005581_70014*) was found in 9% of the 135 strains (12/135), 11 strains belonging to lineage I (21%) and one strain belonging to lineage II (1%) (**Figure III-3**). The genes encoding internalin A, B, C, E, H, J and K, and sigma factor B (*sigB*) were present in all 135 strains (**Figure III-3**). Full length of the *inlA* gene was

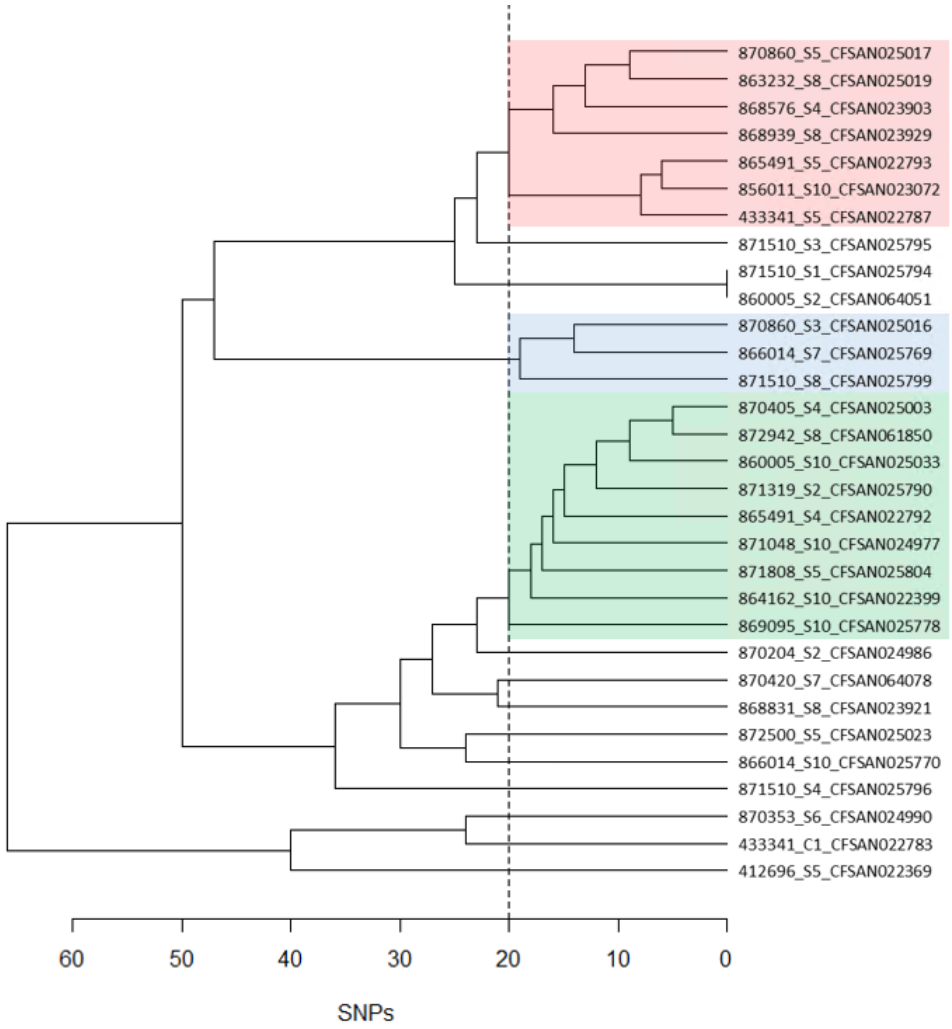


Figure III-4. Phylogenetic tree of the 31 CC14 strains (Cluster D). The single nucleotide polymorphism (SNP) distance matrix was generated using the Center for Food Safety and Applied Nutrition (CFSAN) SNP pipeline. The threshold (20 SNPs) belonging to the same strain is delineated and determined by three types of strains.

observed in all 135 strains (**Figure III-3**). The *inlF* gene was present in 98% of the 135 strains (132/135), 53 strains belonging to lineage I (100%) and 78 strains belonging to lineage II (95%) (**Figure III-3**). The *inlG* gene was found in 59% of the 135 strains (79/135), one strain belonging to lineage I (2%) and 78 strains belonging to lineage II (95%) (**Figure III-3**).

Stress resistance traits among the 135 *L. monocytogenes* strains.

The presence of significant genes responsible for stress resistance of *L. monocytogenes* was determined in the 135 *L. monocytogenes* strains (**Figure III-4**). The genes involved in *L. monocytogenes* biofilm formation, such as *lmo0673, lmo2504, recO* and *luxS* were found in all 135 strains. Other genes involved in biofilm formation, *bapL* and *inlL*, were not found in the lineage I strains, whereas the *bapL* was found in 45 lineage II strains (55%) and the *inlL* was found in 70 lineage II strains (85%). Additional seven genes (*esaA, lmo0453, lmo0543, lmo1224, uvrA, lmo2563* and *lmo2572*) associated with *L. monocytogenes* biofilm formation at a temperature of 15℃ [84] were present in all 135 strains. One plasmid was identified among the 135 strains. One strain (CFSAN025031) recovered from the avocado sample 860005 harbored a *repA* gene. The NCBI BLAST comparison against publicly available *Listeria* plasmid sequences further determined 16 of 56 contigs of CFSAN025031 as plasmid contigs. A total length of 151 Kbp plasmid was identified and harbored a plasmid-borne benzalkonium chloride (BC) resistance cassette, *bcrABC,* and one cadmium resistance gene cassette, *cadA2C2.* Chromosome-borne BC resistance genes *ladR* and *mdrL* were found in all strains. Other BC resistance

genes such as *qacH, qacA, qacC, emrC* and *emrE* were not found in 135 strains. One CC14 strain belonging to lineage II harbored the *Listeria* genomic island 2 (LGI-2, *cadA4C4* and *arsD1A1R1D2R2A2B1B2*), which was previously identified to associate with heavy metal resistance [86].

Other stress resistance genes involved in *L. monocytogenes* survival under extreme conditions such as low pH, high salt concentration, desiccation and low temperature were investigated in the 135 strains (**Figure III-3**). The stress survival islet 1 (SSI-1, *lmo0444* to *lmo0448*), which is involved in the survival under low pH and high salt concentrations, was present in 43 out of the 135 strains (32%), 17 strains belonging to lineage I (32%) and 26 strains belonging to lineage II (32%) (**Figure III-3**). SSI-2 (*lin0464* and *lin0465*), involved in alkaline and oxidative stress resistance, was not found in any strain. Additional genes involved in tolerance of low pH (*arcABCD, gadD2T2, lmo0796, lmo0913* and *lmo2391*), desiccation (*flgD, flhB, fliM, fliP, fliY* and *motB*), high salt concentration (*gbuABC, opuCA, opuCB, opuCC* and *opuCD*) and cold (*lmo0866, lmo1722, lisK, oppA, yycG* and *cspB*) were present in all strains (**Figure III-3**). One of the cold resistance genes, *cspD,* was found in 75 of the 135 strains (56%), none belonging to lineage I strains and 75 strains belonging to lineage II strains (91%) (**Figure III-3**). Lastly, intrinsic antibiotic resistance genes (*fosX, lmo0441, lmo0919, norB* and *sul*) were present in all strains but other acquired antibiotic resistance genes (*aaaA4, aphA, cat_CHL, dfrD, ermB, penA, qnrB, str, tetM* and *tetS*) were not observed in any strain.

Discussion

Analyzing WGS data of *L. monocytogenes* isolates obtained from the nationwide surveillance on whole fresh avocado allowed us to reveal the clonal diversity of *L. monocytogenes* associated with the commodity and provide better knowledge about the genetic diversity and possible transmission pathways of *L. monocytogenes*. In this chapter, we used the 1,827-cgMLST scheme for molecular subtyping of *L. monocytogenes* and identification of duplicated isolates from the same sample caused by the wet pooling sampling approach. After removing 115 duplicated isolates, the remaining 135 strains were considered representative strains associated with whole fresh avocado skin samples.

Lineage II strains were present in higher number of samples(n=82, 61%) compared to lineage I strains (n=53, 39%) among the 135 strains analyzed. This finding is consistent with previous studies suggesting that lineage II strains may have a higher ability to tolerate and survive under various environmental stress **[6, 8, 87, 88]**. Further investigation of genetic elements in the 135 strains showed that genes associated with stress resistance traits of *L. monocytogenes,* such as biofilm formation, tolerance of desiccation and low temperature, were mainly present among the lineage II strains (**Figure III-3**).

A total of 22 CCs and three singletons were identified among the 135 strains. More than half (n=35, 55%) of the positive avocado samples contained two or more strains per sample, emphasizing the importance of appropriate sampling procedures to yield sufficient recovery of different strains of *L. monocytogenes* for population

diversity studies. Within the same CC, strains differed by ≤ 195 alleles, except CC14 and CC5, which showed unexpectedly higher diversity than typical diversity within a CC [28]. CC5 and CC14 strains differed by ≤ 1,182 alleles and ≤ 1,438 alleles, respectively, and did not form a single cluster on the phylogenetic tree (**Figure III-3**). One ST5 strain belonging to CC5 had a distinct genetic profile and differed from other ST363 strains, indicating that that it might belong to a different clone if WGS was used to define *L. monocytogenes* clones (**Figure III-3A**). Similarly, among the five STs classified into CC14, three STs (ST14, ST206 and ST91) did not form a single cluster and also showed distinct genetic profiles from other two STs (ST365 and ST1301), indicating that they might belong to different clones (**Figure III-3B**). This highlights that the inconsistency between the current designation of clones based on 7-locus MLST and the WGS phylogeny, suggesting the need for using WGS to define *L. monocytogenes* clones [79].

Among 22 CCs and 3 singletons identified in this study, CC14, excluding ST14, ST206, and ST91, was present in the highest number of samples (n=35), mainly from imported avocado samples grown in Mexico (89% of 27 avocado samples). This clone was previously identified in an urban environment in 2002 and in two avocado samples in Mexico in 2015, according to the Institut Pasteur database, suggesting that this clone is likely associated with the environment and avocados produced in Mexico. The clone found in the second highest number of samples was CC392 with 16 strains, mostly from domestically grown avocado samples (85% of 13 avocado samples). This clone was previously found in animals such as sheep or cattle in Spain [89], vegetation, and an avocado sample in Chile in 2014, according to the

Institut Pasteur database. Moreover, CC392 was associated with a clinical sample in Canada in 2007, according to the Institut Pasteur database, indicating the potential risk of listeriosis infection caused by hypovirulent clones.

Clustering analysis on the 135 strains obtained from 64 avocado skin samples identified several clusters of closely related strains (≤ 7 alleles or ≤ 12 alleles and ≤ 20 SNPs), suggesting possible transmission or contamination events of *L. monocytogenes.* Clusters A, B, H and I contained the strains belonging to avocado samples grown in California, and Clusters C, D, E, F, G and J contained the strains belonging to two avocado samples grown in Mexico (**Figure III-3**). No clusters were identified from avocado samples grown in different regions, which suggests that there was no transmission or contamination occurred between imported and domestically grown avocado samples at the retail stores. The available metadata were not sufficient to identify the actual transmission or contamination events. For instance, four avocado samples grown in California and belonging to Cluster B and those avocado samples grown in Mexico had incomplete metadata about the growing farm locations. Among 10 clusters identified in this study, Cluster A contained two avocado samples grown in two farms in California located 20 miles away from each other. Cluster H also contained two avocado samples grown in California and they were produced by the same company in California. Cluster I contained three avocado samples grown in California, two of which were produced by the same company located five miles away from the third company. Clusters E and F contained avocado samples produced by the same Mexican company, likely grown in the same avocado farm. Complete metadata would have led to better investigation of whether the closely related *L.*

monocytogenes strains may have originated from a common source or may have been transmitted between different farms.

The presence of major genes and genetic islands associated with the virulence of *L. monocytogenes* was identified among the 135 strains. LIPI-3 was predominantly found in strains belonging to lineage I (93% of 30 strains harbored LIPI-3) and most of them were from avocado samples grown in California (80% of 20 avocado samples). LIPI-4 was also highly found in strains belonging to lineage I (92% of 12 strains harbored LIPI-4) and most of them were from avocado samples grown in California (82% of 11 avocado samples). Interestingly, a full length of *inlA* gene was found in all 135 strains, which is in contrast to previous genetic diversity studies that have reported a wide range of proportions of strains containing PMSCs in *inlA* gene ranging from 2 to 35% **[46, 90, 91]**. Furthermore, all 135 strains harbored major virulence-associated genes, including *sigB* and internalins genes (i.e., *inlABCEHJK*), suggesting that *L. monocytogenes* strains present in the avocado skin samples may carry a relatively high virulence potential.

The prevalence of *L. monocytogenes* on the surface of fresh whole avocados was higher (18%) than other food products screened during FDA prevalence surveys such as raw milk cheese (0.62%, 10/1,606 samples) [92] and sprouts (1.28%, 6/469 samples) [93]. Assessing the WGS data for the biodiversity of *L. monocytogenes* strains recovered from avocado skin samples suggests that the surface of whole fresh avocados may serve as a food vehicle for diverse *L. monocytogenes* clones including hypervirulent and hypovirulent clones. The use of comparative WGS analyses including cgMLST scheme on *L. monocytogenes* strains allowed us to better

understand the possible heterogenicity in virulence and stress resistance of this bacterium. It also highlights the practical and powerful discriminatory approach of cgMLST scheme in revealing the genetically relatedness of *L. monocytogenes,* which can aid in tracing the transmission and contamination source for future outbreak or sporadic case investigations.

Chapter IV: Prevalence and genetic diversity of *L. monocytogenes* in ice cream production facilities

Abstract

In 2015, ice cream products were first recognized as possible food vehicles for *L. monocytogenes*, a psychrophilic bacterium capable of surviving under freezing storage conditions. In the same year, another listeriosis outbreak was recognized in Washington State due to the contamination of ice cream mixes and milkshake maker. The following year, in 2016, the U.S. FDA conducted a nationwide inspection, investigating 89 local ice cream production facilities to assess the prevalence of *L. monocytogenes* by collecting 5,295 environmental samples. In summary, *L. monocytogenes* was detected in 65 samples collected from 19 facilities. After removing duplicates using the cgMLST and SNP analyses, 33 strains were selected for further analyses. Among the 33 strains, two lineages were identified: lineage I (n=22, 67%) and lineage II (n=11, 33%). *In silico* MLST classified the strains into 12 CCs, consisting of 14 STs. CC5 was the most prevalent clone (n=16, 48.5%) followed by CC155 (n=4, 12.1%), CC321 (n=3, 9.1%) and C11 (n=2, 6.1%). Notably, isolates belonging to CC5 strains were detected from 12 out of 19 facilities (63%), which yielded *L. monocytogenes*-positive samples. Additionally, 28 out of the 33 strains carried plasmids ranging in size from 58 to 161 Kb. The presence of major virulence genes and genes responsible for stress resistance was determined. Furthermore, a comparative analysis for CC5 strains was conducted among the strains identified

during the 2016 FDA surveillance and the *L. monocytogenes* strains associated with the listeriosis outbreak occurred in Washington State in 2015.

Introduction

In early 2015, an outbreak of listeriosis associated with the consumption of contaminated ice cream products was recognized, leading to 10 hospitalizations and three deaths **[94]**. A subsequent study revealed that this outbreak was caused by strains belonging to the ST5, classified into CC5 group **[95]**. Notably, *L monocytogenes* was initially detected in ice cream products at a distribution center during routine sampling efforts **[96]**. Subsequently, patients infected with *L. monocytogenes* were found to have PFGE patterns that matched those previously identified in the ice cream production facilities **[96]**. The rapid recognition of this outbreak was partically due to real-time WGS-based surveillance, which enabled storage of sequence genomes in a publically available database, generating a daily-updated SNP-based WGS tree **[95]**. This incident highlighted the practical and effective application of WGS as a powerful tool for investigating outbreaks. It demonstrated the advantages and unique opportunities provided by WGS in identifying atypical food vehicles for *L. monocytogenes*, which are often considered to pose low risks of causing listeriosis and not supporting the growth of *L. monocytogenes* **[3]**.

In 2016, the U.S. FDA conducted a nationwide inspection involving collecting environmental samples from ice cream production facilities across the country to assess the prevalence of *L. monocytogenes* **[97]**. A total of 89 ice cream facilities located in 32 states were inspected and a total of 5,295 environmental samples were collected and analyzed for *L. monocytogenes* **[97]**. The samples were collected from various locations within the facility, including floors, drains, equipment surfaces, and

other surfaces in production and storage areas **[97]**. Overall, *L. monocytogenes* was detected in 19 of the 89 facilities, with an establishment-based prevalence of 21.3%, and in 65 of the 5,295 samples, with a sample-based prevalence of 1.25% **[97]**. As a part of the investigation, the FDA conducted WGS of all *L. monocytogenes* isolates were sequenced and the WGS data were stored in the GenomeTrakr database.

This chapter focuses on the assessment of WGS data of *L. monocytogenes* isolates associated with ice cream production facilities. The objectives were to 1) determine the prevalence of *L. monocytogenes* genotypes in such environments, 2) investigate the genetic diversity to infer possible transmission patterns of *L. monocytogenes*, and 3) reveal the phylogenetic relationship with the *L. monocytogenes* strains collected from a previous outbreak investigation to assess the persistence of *L. monocytogenes* in ice cream production facilities.

Materials and Methods

Data collection of *L. monocytogenes* isolates

In this chapter, WGS data of 65 *L. monocytogenes* isolates, which were obtained from environmental samples collected from 19 ice cream production processing facilities in 2016 and 2017, were retrieved from the GenomeTrakr database. For a comparative analysis of isolates belonging to CC5 strains, 33 *L. monocytogenes* isolates previously analyzed during the outbreak investigation occurred in Washington State were included **[98]**. The paired-end raw reads were *de novo* assembled using SKESA v0.24 **[55]** with default parameters. The quality of genomes was assessed using QUAST v5.0.2 **[77]** following the same threshold described in Chapter III.

Core-genome MLST and determination of duplicated isolates

The 65 *L. monocytogenes* genomes were subjected to the 1,827-cgMLST scheme. Following the thresholds described in Chapter 3, duplicated isolates were identified and subsequently eliminated. As a result, 33 *L. monocytogenes* strains were selected (**Table IV-1**) for further analyses as described in Chapter III.

Determination of lineages, PCR-serogroups, CCs and STs and phylogenetic analysis

The determination of lineages, PCR-serogroups, CCs and STs for the 33 *L. monocytogenes* strains was performed as described in Chapter III. A phylogenetic tree

was built using a single-linkage algorithm based on the pairwise AD profile of the cgMLST scheme in Ridom SeqSphere^{+}.

Table IV-1. List of 33 *L. monocytogenes* strains recovered from 19 ice cream production facilities and analyzed in this chapter.

Sample ID	FDA Accession	State	Facility	Lineage	Serogroup	CC [a]	ST [b]	Plasmid (Kb)
923816	CFSAN065191	MD	A	I	IIb	CC5	1812	60
952966	CFSAN055749	WA	B	I	IIb	CC5	5	131
960927	CFSAN059983	CA	C	I	IIb	CC5	5	61
964468	CFSAN056968	NY	D	I	IIb	CC5	5	72
964468	CFSAN056971	NY	D	I	IIb	CC5	1812	71
968336	CFSAN060656	IN	E	I	IIb	CC5	5	161
968336	CFSAN060658	IN	E	I	IIb	CC5	5	73
968336	CFSAN060667	IN	E	I	IIb	CC5	5	160
968615	CFSAN058489	MI	F	II	IIa	CC155	155	91
968617	CFSAN058488	MI	F	II	IIa	CC14	14	91
968833	CFSAN063816	PA	G	I	IIb	CC5	1675	60
972048	CFSAN063069	ME	H	I	IIb	CC5	5	NA[c]
973150	CFSAN069609	FL	I	I	IIb	CC5	5	61
973150	CFSAN069613	FL	I	I	IIb	CC224	224	82
980730	CFSAN062987	UT	J	II	IIa	CC321	321	63
980730	CFSAN062988	UT	J	II	IIa	CC321	321	64
980730	CFSAN062989	UT	J	II	IIa	CC193	796	NA
988150	CFSAN064770	NJ	K	I	IIb	CC5	5	58
988528	CFSAN065801	CA	L	I	IIb	CC5	5	113
993173	CFSAN062612	CA	M	II	IIa	CC11	11	NA
995411	CFSAN067787	OH	N	II	IIa	CC11	11	91
995411	CFSAN067788	OH	N	I	IVb	CC2	2	150
995411	CFSAN067789	OH	N	I	IIb	CC5	5	149
999084	CFSAN065742	CT	O	II	IIa	CC321	321	65

999084	CFSAN065744	CT	O	I	IVb	CC1	308	NA
1002094	CFSAN067673	GA	P	I	IVb	CC6	6	90
1004600	CFSAN064785	FL	Q	I	IIb	CC88	296	82
1009994	CFSAN067666	MD	R	I	IIb	CC5	5	151
1009994	CFSAN067667	MD	R	I	IIb	CC5	5	91
1014164	CFSAN069118	MA	S	I	IIb	CC1041	1041	NA
1014164	CFSAN069121	MA	S	II	IIa	CC155	155	82
1014164	CFSAN069122	MA	S	II	IIa	CC155	155	87
1014164	CFSAN069124	MA	S	II	IIa	CC155	155	83

[a] Composite (C) and subsample (S).

[b] Sequence type (ST).

Table IV-2. List of 33 *L. monocytogenes* strains collected during the listeriosis outbreak investigation in Washington State in 2015 (Li et al., 2017).

Isolate ID	Collection time	Source Type	SRA ID
PNUSAL001207	November, 2014	Clinical	SRR1745448
PNUSAL001241	December, 2014	Clinical	SRR1745474
CFSAN028842	December, 2014	Ice cream/Hospital X	SRR3130313
CFSAN028843	December, 2014	Ice cream/Hospital X	SRR3091402
CFSAN028844	December, 2014	Ice cream/Hospital X	SRR3130327
CFSAN028845	December, 2014	Ice cream/Hospital X	SRR3066080
CFSAN028846	December, 2014	Ice cream/Hospital X	SRR3130329
CFSAN028847	December, 2014	Ice cream/Hospital X	SRR3091403
CFSAN028848	December, 2014	Ice cream/Company A	SRR3091404
CFSAN028849	December, 2014	Ice cream/Company A	SRR3091405
CFSAN028850	December, 2014	Ice cream/Company A	SRR3130331
CFSAN028851	December, 2014	Ice cream/Company A	SRR3091406
CFSAN028852	December, 2014	Environmental/Company A	SRR3130333
CFSAN028853	December, 2014	Environmental/Company A	SRR3130335
CFSAN028855	December, 2014	Environmental/Company A	SRR3130404
CFSAN028856	December, 2014	Environmental/Company A	SRR3130406
CFSAN028857	December, 2014	Environmental/Company A	SRR3130409
CFSAN028858	December, 2014	Environmental/Company A	SRR3130413
CFSAN028859	December, 2014	Environmental/Company A	SRR3130415
CFSAN028860	December, 2014	Environmental/Company A	SRR3130350
CFSAN028861	December, 2014	Environmental/Company A	SRR3130375
CFSAN029502	December, 2014	Environmental/Company A	SRR3130341
CFSAN030692	March, 2015	Environmental/Company A	SRR1974103
CFSAN032836	April, 2015	Environmental/Company B	SRR2035442
CFSAN043359	November, 2015	Ice cream/Hospital X	SRR3052035
CFSAN043360	November, 2015	Ice cream/Hospital X	SRR3053137
CFSAN043361	November, 2015	Ice cream/Hospital X	SRR3086932
CFSAN043362	November, 2015	Environmental/Hospital X	SRR3086935
CFSAN043363	November, 2015	Environmental/Hospital X	SRR3086936
CFSAN043364	November, 2015	Environmental/Hospital X	SRR3052036
PNUSAL001911	November, 2015	Clinical	SRR2994642
CFSAN028854	December, 2014	Environmental/Company A	SRR3130337
CFSAN004336	NA	Food	SRR1818032

Identification of plasmids

The identification of plasmids was conducted using PLACNETw, a web-based tool **[99]**. Paired reads were uploaded, and plasmid sequences were detected based on the BLAST built in the PLACNETw against the NCBI database. Among the 33 *L. monocytogenes* strains, 28 strains were determined to harbor plasmid sequences and these regions were extracted via PLACNETw. In addition, the presence of *repA* sequences was determined using four different types of *repA* sequences described by Kuenne et al. and Schmitz-Esser et al. **[85, 100]** and confirmed that five strains did not contain any plasmids.

Results

Description of the selected *L. monocytogenes* isolates

A total of 65 *L. monocytogenes* isolates were recovered from 65 environmental samples collected from 19 ice cream production facilities **[97]**. Using the 1,827-cgMLST scheme for pairwise analysis, 28 isolates recovered from the same sample were identified to likely belong to the same strain with a threshold of 7 AD, and thus were considered as duplicates. Additionally, four isolates were determined as duplicates from the same sample with ≤ 12 AD and ≤ 20 SNPs. After removing the duplicates, 33 strains were selected to represent the 19 facilities (**Table IV-1**). Among the 19 facilities, 10 facilities yielded one *L. monocytogenes* strain, five yielded two strains, and three yielded three strains. One facility located in MA (Facility S) contained four strains which was the highest number of strains in a faiclity (**Figure IV-1**).

PCR-serogroups and phylogenetic analysis

The phylogenetic tree of the 33 strains revealed two distinct evolutionary lineages. Among these strains, 11 strains (67%) belonged to lineage I and 22 strains (33%) belonged to lineage II. The pairwise AD among lineage I strains ranged from 12 to 1,296, while for lineage II strains, it ranged from 8 to 1,459. These two lineages differed by 1,763 to 1,790 AD. Furthermore, three major serogroups were identified, with the serogroup IIb (1/2b, 3b and 7, n=19) being the most prevalent, followed by serogroup IIa (1/2a and 3a, n=11) and serogroup IVb (4b, 4d and 4e, n=3) (**Figure**

IV-2).

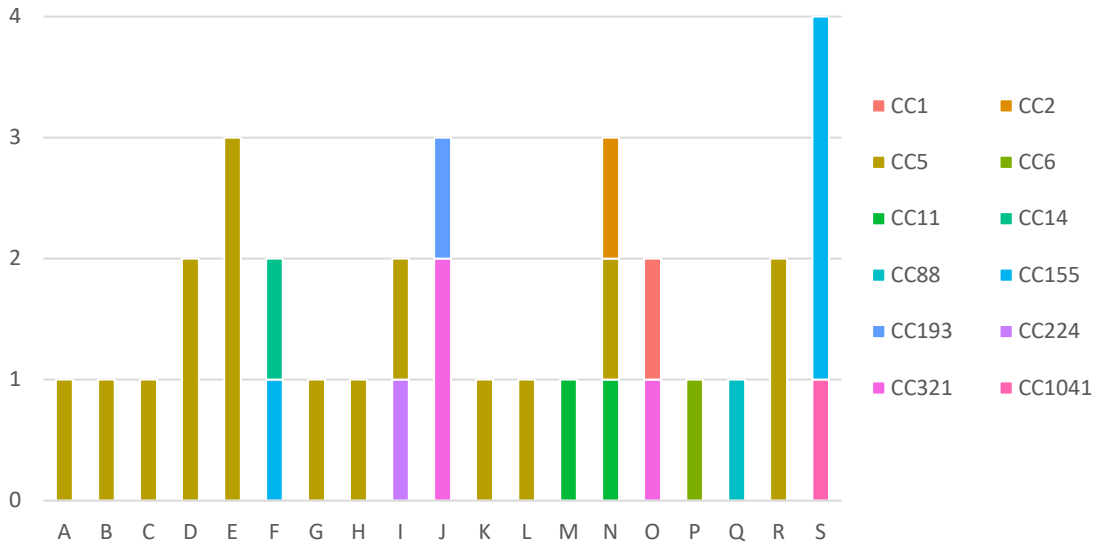


Figure IV-1. Prevalence of the 33 *L. monocytogenes* strains among the 19 ice cream production facilities (A to S). The number of isolates per facility is shown in different colors indicating the 12 clonal complexes (CCs) identified.

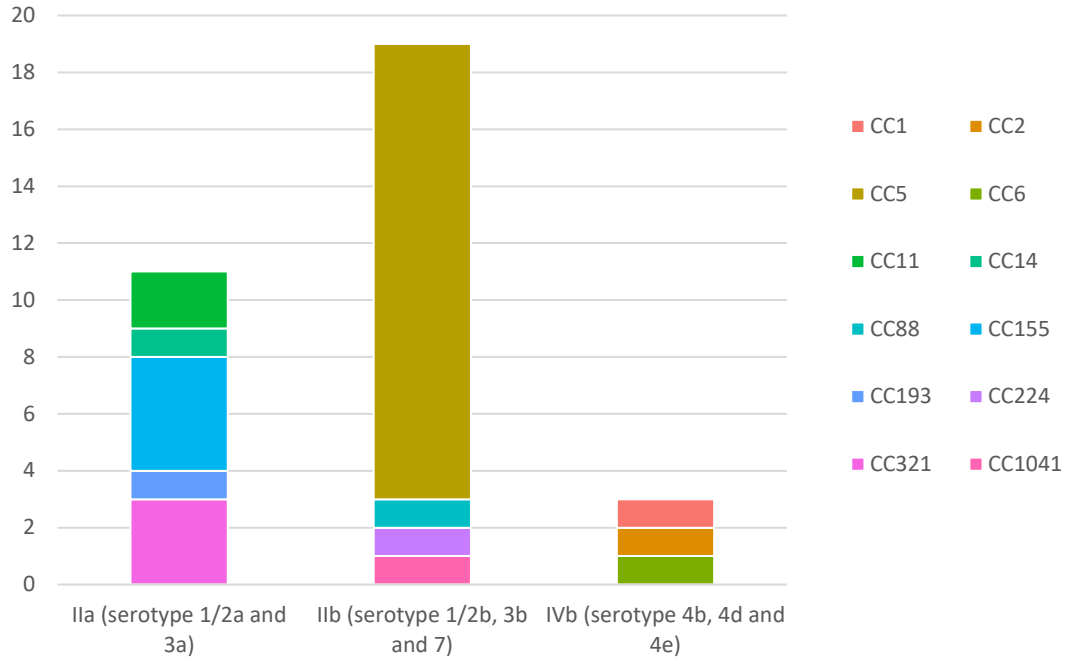


Figure IV-2. Distribution of PCR-serogroups among the 33 *L. monocytogenes* strains. The number of isolates per serogroup is shown in different colors indicating the 12 clonal complexes (CCs) identified.

Among the 33 strains, a total of 12 CCs, comprising 14 STs, were identified (**Table IV-1**). Each CC formed a monophyletic clade on the phylogenetic tree constructed based on the 1,827-cgMLST (**Figure IV-3**). Significantly, CC5, which has been implicated in the 2015 listeriosis outbreak linked to ice cream products, was the most prevalent clone. Out of the 33 isolates, 16 isolates (48%) were identified as CC5 and differed by 12 to 157 AD. Notably, these CC5 isolates were found in 12 facilities (63% of the 19 facilities) located across 10 different states (**Table IV-1 and Figure IV-1**). Following CC5, CC115 contianed four isolates that differed by 8 to 41 AD, and CC321 contained three isolates differed by 11 to 12 AD (**Figure IV-3**).

Isolates obtained from different samples and exhibiting ≤ 12 AD were subjected to the CFSAN SNP pipeline to identify genomic SNP variations. A cluster was initially identified, comprising three isolates belonging to CC321, with ≤ 12 AD (**Figure IV-3**). However, further analysis based on the SNP matrix revealed that these isolates were indeed different strains, as they differed by > 25 SNPs.

Genetic traits of the 33 *L. monocytogenes* strains

To investigate the genetic diversity among the 33 *L. monocytogenes* strains, the genes associated with virulence potential and stress response were examined. All isolates contained LIPI-1, as expected. Among the 33 strains, four strains (12%) belonging to CC224, CC1041, CC6 and CC1, were found to contain LIPI-3, while one isolate (3%) belonging to CC88 were found to contain LIPI-4. The genes encoding internalin A, B, C, E, H, J and K were present in all 33 strains. The *inlF* gene was present in 33 strains, excluding three strains belonging to CC11 and CC14. The *inlG* gene was found in eight strains belonging to CC6, CC155 and CC321. The

PMSCs in *inlA* gene were found in eight strains belonging to CC5 (n=3), CC224 (n=1), CC321 (n=3) and CC193 (n=1).

Genes responsible for stress response in *L. monocytogenes* were investigated. Specifically, the genes known to be involved in *L. monocytogenes* biofilm formation, such as *lmo0673, lmo2504, recO* and *luxS* were found in all 33 strains. Additionally, the gene *bapL,* was found in six strains (18%) belonging to CC321, CC11 and CC14 and another gene *inlL,* was detected in eight strains (24%) belonging to CC155, CC321 and CC193. Additional seven genes (*esaA, lmo0453, lmo0543, lmo1224, uvrA, lmo2563* and *lmo2572*) associated with *L. monocytogenes* biofilm formation at relatively low temperature (15℃) **[84]** were present in all 33 strains. Furthermore, 28 out of 33 strains carried plasmids (85%) ranging in size from 58 to 161 Kb. These plasmids contained gene cassettes associated with BC and heavy metal resistance. Specifically, *bcrABC* were found in 25 strains (76%), *cadA1C1* were found in 10 strains (30%) and *cadA2C2* were found in 19 strains (58%). Moreover, all strains possessed chromosome-borne BC resistance genes *ladR* and *mdrL.* Other BC resistance genes such as *qacA, qacC, emrC* and *emrE* were not found in 33 strains, except for the *qacH* gene, which was found in one strain belonging to CC5. Furthermore, LGI-2, a genomic island including *cadA4C4* and *arsD1A1R1D2R2A2B1B2* was found in three strains belonging to CC1, CC2 and CC14. The survival-associated gene SSI-1, which aids in tolerating low pH and high salt concentrations, was present in 27 strains (82%). However, SSI-2, which is involved in resistance against alkaline and oxidative stress, was absent in all strains. All strains carried intrinsic antibiotic resistance genes (*fosX, lmo0441, lmo0919, norB*

and *sul*), while no acquired antibiotic resistance genes (*aaaA4, aphA, cat_CHL, dfrD, ermB, penA, qnrB, str, tetM* and *tetS*) were detected, except for the presence of the *aaaA4* gene in one strain belonging to CC1041.

Phylogenetic analysis for CC5 strains

A comparison was performed between the isolates in the 2016 surveillance and strains involved in previous outbreak investigations involving ice cream. Among these incidents, one CC5 strain, not linked to clinical cases, but collected during the outbreak investigation in Washington State, showed a genetic match with 16 CC5 strains collected during the 2016 surveillance. These strains were subjected to the CFSAN SNP pipeline, and a matrix of SNPs was generated for constructing a phylogenetic tree (**Figure IV-3**). Notably, the 16 CC5 strains and the CC5 strain recovered from one of the environmental samples obtained from Company A **[98]** differed by ≤ 23 SNPs.

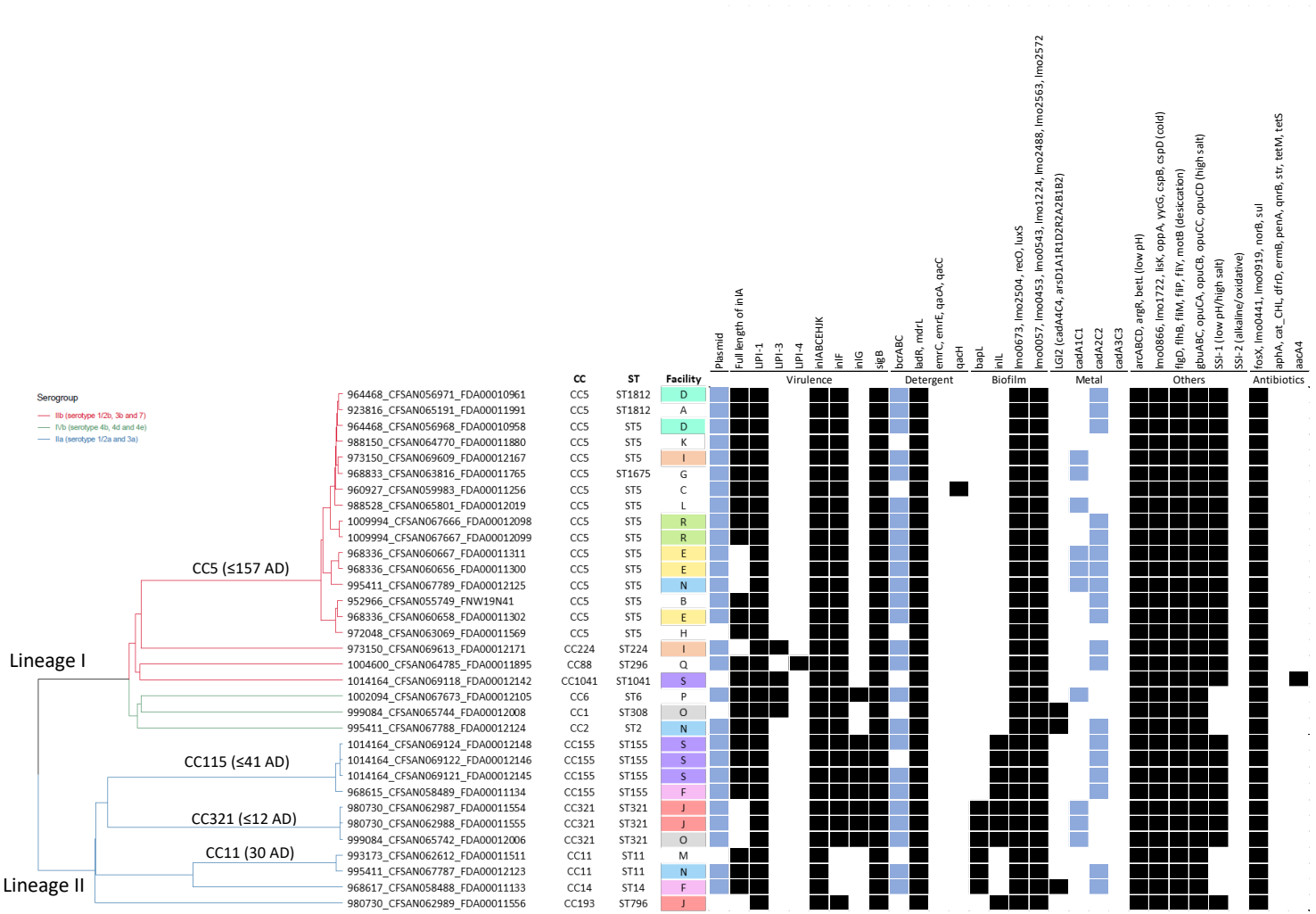


Figure IV-3. Phylogenetic tree of the 33 *L. monocytogenes* strains with the presence (black or blue) or absence (blank) of genes.

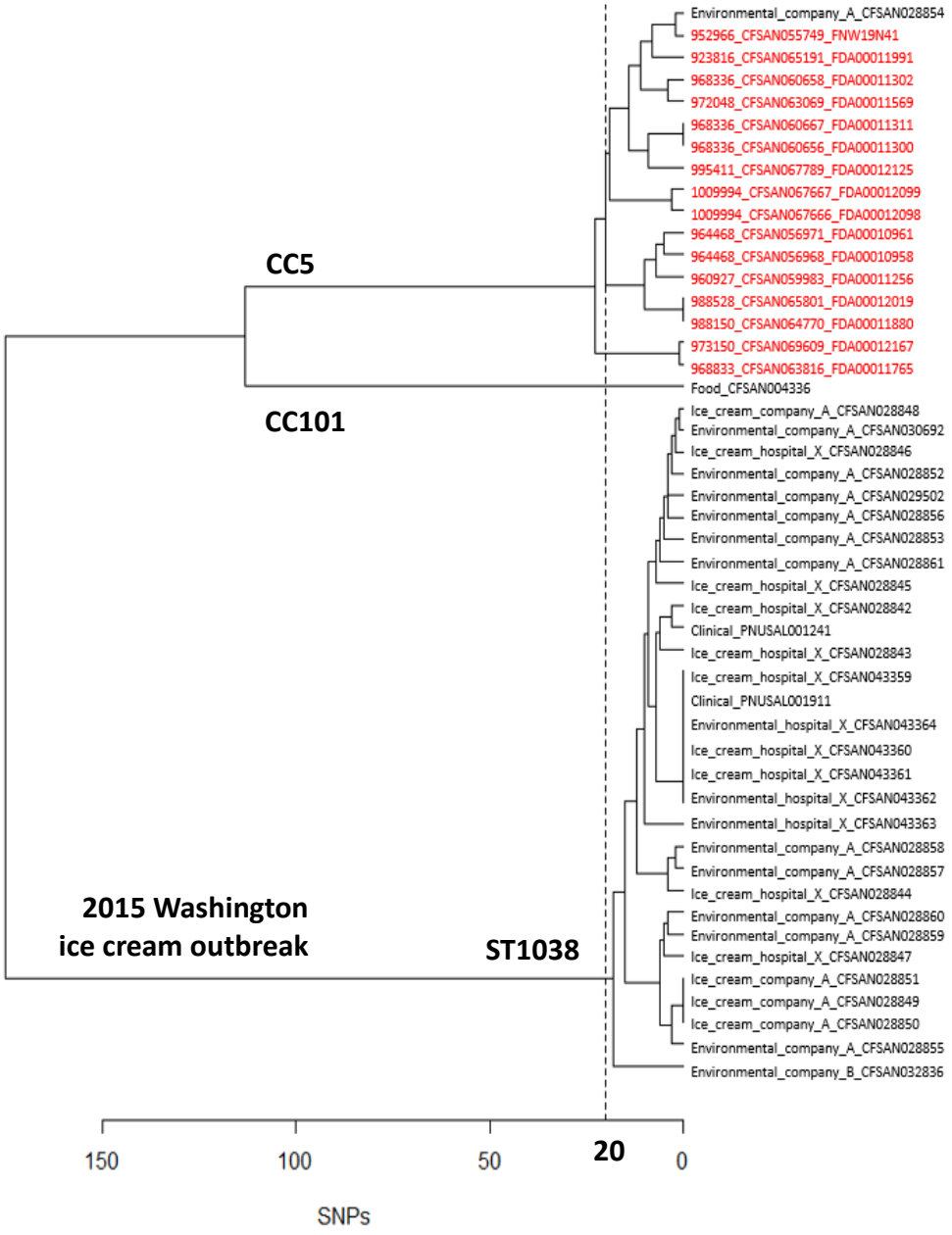


Figure IV-4. Phylogenetic tree based on the distance matrix generated using CFSAN SNP pipeline. The 16 *L. monocytogenes* strains recovered from the 2016 FDA surveillance are indicated in red.

Discussion

Among the 89 ice cream production facilities investigated, environmental samples collected from 19 facilities yielded positive for *L. monocytogenes*, resulting in a prevalence of 21.3% **[97]**. A total of 65 *L. monocytogenes* isolates were detected and sequenced during this surveillance **[97]**. After removing duplicated isolates collected from the same environmental sample, 33 strains were selected for further analyses. A total of 12 CCs of *L. monocytogenes* were identified among the 33 strains. Each CC, containing more than two isolates, differed by ≤ 157 AD. The most prevalent clone was CC5, which was previously associated with the multistate listeriosis outbreak linked to ice cream products in 2015 **[101]**. CC5 strains were determined in 12 out of 19 facilities, representing 63.2%, located in 10 different states across the country, suggesting a potential common supplier transmitting *L. monocytogenes* to these facilities. The most prevalent serogroup was IIb (1/2b, 3b and 7, 58%), followed by serogroup IIa (1/2a and 3a, 33%) and serogroup IV (4b, 4d and 4e, 9%).

The genetic diversity among the 33 *L. monocytogenes* strains was investigated. A total of 25 strains (76%) contained a full-length of *inlA* gene, which is higher compared to previous studies on genetic diversity that have reported a wide range of proportions of strains containing PMSCs in the *inlA* gene, ranging from 2% to 35% **[46, 90, 91]**. LIPI-3 was found in four strains, while LIPI-4 was only found in one strain belonging to CC88. Furthermore, a total of 28 out of 33 strains carried plasmids harboring genes responsible for BC tolerance and metal resistance. These findings provide insights into how these strains were able to survive and potentially

persist in ice cream production facilities. However, due to the lack of complete genome sequences, the comprehensive comparison of horizontal gene transfer or characterization of these plasmid sequences were challenging.

Phylogenetic analysis performed on the *L. monocytogenes* CC5 strains and previously collected strains from the listeriosis outbreak investigation in Washington State revealed a close genetic relationship between an environmental strain collected from Company A in December 2014 and an environmental strain recovered from Facility B (2 SNPs), which was collected after August 2016. Notably, these two facilities belong to the same company in Washington State, suggesting that this strain could have been persisted over 20 months. Furthermore, other CC5 strains were all closely related (≤ 23 SNPs), suggesting the possibility of a common source that could potentially facilitate the transmission of *L. monocytogenes* to these facilities. No other matches were determined among strains collected from the 19 facilities, and therefore, WGS data are not inferring any recent contamination events.

CC5 strains have been implicated in multiple listeriosis outbreaks in the U.S., linked to cantaloupe, cheese, stone fruit, and ice cream products [6-8]. It is noteworthy that L. monocytogenes is frequently isolated in unpasteurized dairy cow milk, as well as dairy products such as soft cheeses, pasteurized milk that has been contaminated, feces of both asymptomatic dairy cows and cows with listeriosis, and dairy farm environments [9]. Given the widespread occurrence of CC5 strains in the ice cream processing facilities throughout the country, it is plausible that the prevalence of CC5 in ice cream products could be derived from the prevalence of CC5 in dairy ingredients used during ice cream production. To test this hypothesis, an

investigation of clonal diversity of the dairy food and ingredients can be performed. Another possible explanation is that CC5 could have an enhanced ability to survive under low temperature and wet conditions of ice cream production facilities [8] which allowed them to persist for an extended period of time. Phenotype comparison between CC5 and other *L. monocytogenes* genotypes could provide more insights into the fit of CC5 in ice cream and ice cream production environment.

Chapter V: Summary and future directions

Understanding population structure and genetic diversity of *L. monocytogenes* helps us to investigate bacterial evolution, track its spread and transmission, identify ecological niches of genotypes, identify high-risk strains, and establish nomenclature for easier information exchange with previous and future studies.

Chapter 1 focused on identifying and characterizing the MGEs associated with *L. monocytogenes* recovered from meat and poultry processing facilities. This was achieved by utilizing both long-read sequencing and short-read sequencing technologies to generate complete genomes. The objective was to investigate the potential persistence of CC6 strains showing outbreak PFGE patterns over several years and determine their short-term evolution by examining the acquisition of unique genetic regions through horizontal gene transfer of MGEs. These unique genetic regions also serve as genetic markers for tracking the transmission of *L. monocytogenes* strains in these facilities.

Chapter 2 aimed to assess the prevalence of genotypes and genetic diversity of *L. monocytogenes* strains recovered from whole fresh avocados. The study was motivated by the unexpectedly high prevalence of *L. monocytogenes* observed during the 2014 FDA surveillance. The objective was to identify genotypes that may be unique to growing regions and to understand possible transmission pathways of *L. monocytogenes* strains in the production and retail environment of avocados. Notably, no possible transmission events were observed at the retail sites, but contamination and transmission of *L. monocytogenes* at the growing regions was possible.

In Chapter 3, *L. monocytogenes* strains recovered from ice cream production facilities in the U.S. were analyzed. This chapter aimed to determine the prevalence of genotypes and genetic diversity of *L. monocytogenes* to understand their genetic characteristics and possible transmission pathways between the facilities. Strains collected from facilities during the outbreak investigation in Washington State in 2015 were included for comparative analysis to assess the persistence of *L. monocytogenes*.

Together, these chapters provide valuable insights into the persistence, possible transmission pathways, and genetic characteristics of *L. monocytogenes* in select foods and food processing environments. However, further investigations are needed to elucidate the factors or transmission pathways that contribute to the high prevalence of specific genotypes in certain foods or food processing environments. Such investigations can help inform the development of preventive and control strategies to mitigate *L. monocytogenes* contamination. Moreover, the generation of complete genomes can aid in the identification and characterization of novel genetic markers. In addition, the collection and availability of more comprehensive metadata associated with the isolates would greatly facilitate the interpretation and applicability of the results, further strengthening the impact of the research on pathogen surveillance and management practices. Identification of clones unique to certain food, food growing regions and food manufacturing facilities could lead to investigation of phenotypes that could facilitate the survival of L. monocytogenes in these foods and environments.

Bibliography

1. Swaminathan, B. and P. Gerner-Smidt, *The epidemiology of human listeriosis.* Microbes and infection, 2007. **9**(10): p. 1236-1243.
2. Hamon, M., H. Bierne, and P. Cossart, *Listeria monocytogenes: a multifaceted model.* Nature Reviews Microbiology, 2006. **4**(6): p. 423-434.
3. Buchanan, R.L., et al., *A review of Listeria monocytogenes: An update on outbreaks, virulence, dose-response, ecology, and risk assessments.* Food Control, 2017. **75**: p. 1-13.
4. Carpentier, B. and O. Cerf, *Persistence of Listeria monocytogenes in food industry equipment and premises.* International journal of food microbiology, 2011. **145**(1): p. 1-8.
5. Chenal-Francisque, V., et al., *Worldwide distribution of major clones of Listeria monocytogenes.* Emerging infectious diseases, 2011. **17**(6): p. 1110.
6. Orsi, R.H., H.C. den Bakker, and M. Wiedmann, *Listeria monocytogenes lineages: genomics, evolution, ecology, and phenotypic characteristics.* International Journal of Medical Microbiology, 2011. **301**(2): p. 79-96.
7. Hyden, P., et al., *Whole genome sequence-based serogrouping of Listeria monocytogenes isolates.* Journal of Biotechnology, 2016. **235**: p. 181-186.
8. Haase, J.K., et al., *The ubiquitous nature of Listeria monocytogenes clones: a large-scale Multilocus Sequence Typing study.* Environmental microbiology, 2014. **16**(2): p. 405-416.
9. Maury, M.M., et al., *Uncovering Listeria monocytogenes hypervirulence by harnessing its biodiversity.* Nature genetics, 2016. **48**(3): p. 308.
10. Cotter, P.D., et al., *Listeriolysin S, a novel peptide haemolysin associated with a subset of lineage I Listeria monocytogenes.* PLoS Pathog, 2008. **4**(9): p. e1000144.
11. Lecuit, M., et al., *A transgenic model for listeriosis: role of internalin in crossing the intestinal barrier.* Science, 2001. **292**(5522): p. 1722-1725.
12. Nightingale, K., et al., *Select Listeria monocytogenes subtypes commonly found in foods carry distinct nonsense mutations in inlA, leading to expression of truncated and secreted internalin A, and are associated with a reduced invasion phenotype for human intestinal epithelial cells.* Applied and Environmental Microbiology, 2005. **71**(12): p. 8764-8772.
13. Meinersmann, R.J., et al., *Multilocus sequence typing of Listeria monocytogenes by use of hypervariable genes reveals clonal and recombination histories of three lineages.* Applied and environmental microbiology, 2004. **70**(4): p. 2193-2203.
14. Swaminathan, B., et al., *PulseNet: the molecular subtyping network for foodborne bacterial disease surveillance, United States.* Emerging infectious diseases, 2001. **7**(3): p. 382.
15. Lomonaco, S. and D. Nucera, *Molecular subtyping methods for Listeria monocytogenes: Tools for tracking and control.* DNA Methods in Food

Safety: Molecular Typing of Foodborne and Waterborne Bacterial Pathogens, 2014: p. 303-336.
16. Liu, D., *Identification, subtyping and virulence determination of Listeria monocytogenes, an important foodborne pathogen.* Journal of medical microbiology, 2006. **55**(6): p. 645-659.
17. Nyarko, E.B. and C.W. Donnelly, *Listeria monocytogenes: strain heterogeneity, methods, and challenges of subtyping.* Journal of food science, 2015. **80**(12): p. M2868-M2878.
18. Wiedmann, M., *Molecular subtyping methods for Listeria monocytogenes.* Journal of AOAC International, 2002. **85**(2): p. 524-532.
19. Borucki, M.K. and D.R. Call, *Listeria monocytogenes serotype identification by PCR.* Journal of Clinical Microbiology, 2003. **41**(12): p. 5537-5540.
20. Doumith, M., et al., *Differentiation of the major Listeria monocytogenes serovars by multiplex PCR.* Journal of clinical microbiology, 2004. **42**(8): p. 3819-3822.
21. Salcedo, C., et al., *Development of a multilocus sequence typing method for analysis of Listeria monocytogenes clones.* Journal of clinical microbiology, 2003. **41**(2): p. 757-762.
22. Nightingale, K., *Listeria monocytogenes: knowledge gained through DNA sequence-based subtyping, implications, and future considerations.* Journal of AOAC International, 2010. **93**(4): p. 1275-1286.
23. Stessl, B., I. Rückerl, and M. Wagner, *Multilocus sequence typing (MLST) of Listeria monocytogenes.* Listeria monocytogenes: Methods and Protocols, 2014: p. 73-83.
24. Graves, L.M. and B. Swaminathan, *PulseNet standardized protocol for subtyping Listeria monocytogenes by macrorestriction and pulsed-field gel electrophoresis.* International journal of food microbiology, 2001. **65**(1-2): p. 55-62.
25. Graves, L.M., et al., *Microbiological aspects of the investigation that traced the 1998 outbreak of listeriosis in the United States to contaminated hot dogs and establishment of molecular subtyping-based surveillance for Listeria monocytogenes in the PulseNet network.* Journal of Clinical Microbiology, 2005. **43**(5): p. 2350-2355.
26. Stasiewicz, M.J., et al., *Whole-genome sequencing allows for improved identification of persistent Listeria monocytogenes in food-associated environments.* Applied and environmental microbiology, 2015. **81**(17): p. 6024-6037.
27. Wieczorek, K., A. Bomba, and J. Osek, *Whole-genome sequencing-based characterization of Listeria monocytogenes from fish and fish production environments in Poland.* International Journal of Molecular Sciences, 2020. **21**(24): p. 9419.
28. Chen, Y., et al., *Core genome multilocus sequence typing for identification of globally distributed clonal groups and differentiation of outbreak strains of Listeria monocytogenes.* Applied and environmental microbiology, 2016. **82**(20): p. 6258-6272.

29. Centers for Disease Control Prevention. *Listeria (Listeriosis)*. 2021; Available from: https://www.cdc.gov/listeria/index.html.
30. Fleming, D.W., et al., *Pasteurized milk as a vehicle of infection in an outbreak of listeriosis.* New England journal of medicine, 1985. **312**(7): p. 404-407.
31. Kase, J.A., G. Zhang, and Y. Chen, *Recent foodborne outbreaks in the United States linked to atypical vehicles—lessons learned.* Current Opinion in Food Science, 2017. **18**: p. 56-63.
32. Zhu, M., et al., *Control of Listeria monocytogenes contamination in ready-to-eat meat products.* Comprehensive Reviews in Food Science and Food Safety, 2005. **4**(2): p. 34-42.
33. Mead, P., et al., *Nationwide outbreak of listeriosis due to contaminated meat.* Epidemiology & Infection, 2006. **134**(4): p. 744-751.
34. Gottlieb, S.L., et al., *Multistate outbreak of listeriosis linked to turkey deli meat and subsequent changes in US regulatory policy.* Clinical infectious diseases, 2006. **42**(1): p. 29-36.
35. Chen, Y., W. Zhang, and S.J. Knabel, *Multi-virulence-locus sequence typing clarifies epidemiology of recent listeriosis outbreaks in the United States.* Journal of clinical microbiology, 2005. **43**(10): p. 5291-5294.
36. Smith, A.M., et al., *Outbreak of Listeria monocytogenes in South Africa, 2017–2018: Laboratory activities and experiences associated with whole-genome sequencing analysis of isolates.* Foodborne pathogens and disease, 2019. **16**(7): p. 524-530.
37. Verghese, B., et al., *comK prophage junction fragments as markers for Listeria monocytogenes genotypes unique to individual meat and poultry processing plants and a model for rapid niche-specific adaptation, biofilm formation, and persistence.* Applied and environmental microbiology, 2011. **77**(10): p. 3279-3292.
38. Chen, Y. and S.J. Knabel, *Prophages in Listeria monocytogenes contain single-nucleotide polymorphisms that differentiate outbreak clones within epidemic clones.* Journal of clinical microbiology, 2008. **46**(4): p. 1478-1484.
39. Yang, H., et al., *Microevolution and gain or loss of mobile genetic elements of outbreak-related Listeria monocytogenes in food processing environments identified by whole genome sequencing analysis.* Frontiers in microbiology, 2020. **11**: p. 866.
40. Nelson, K.E., et al., *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 research, 2004. **32**(8): p. 2386-2395.
41. Dutta, V., et al., *Genetic characterization of plasmid-associated triphenylmethane reductase in Listeria monocytogenes.* Applied and Environmental Microbiology, 2014. **80**(17): p. 5379-5385.
42. Elhanafi, D., V. Dutta, and S. Kathariou, *Genetic characterization of plasmid-associated benzalkonium chloride resistance determinants in a Listeria monocytogenes strain from the 1998-1999 outbreak.* Applied and environmental microbiology, 2010. **76**(24): p. 8231-8238.

43. Aase, B., et al., *Occurrence of and a possible mechanism for resistance to a quaternary ammonium compound in Listeria monocytogenes.* International journal of food microbiology, 2000. **62**(1-2): p. 57-63.
44. Ochman, H., J.G. Lawrence, and E.A. Groisman, *Lateral gene transfer and the nature of bacterial innovation.* nature, 2000. **405**(6784): p. 299-304.
45. Moura, A., et al., *Whole genome-based population biology and epidemiological surveillance of Listeria monocytogenes.* Nature microbiology, 2016. **2**(2): p. 1-10.
46. Chen, Y., et al., *Genetic diversity and profiles of genes associated with virulence and stress resistance among isolates from the 2010-2013 interagency Listeria monocytogenes market basket survey.* PLoS One, 2020. **15**(4): p. e0231393.
47. Dorscht, J., et al., *Comparative genome analysis of Listeria bacteriophages reveals extensive mosaicism, programmed translational frameshifting, and a novel prophage insertion site.* Journal of bacteriology, 2009. **191**(23): p. 7206-7215.
48. Hingston, P., et al., *Genotypes associated with Listeria monocytogenes isolates displaying impaired or enhanced tolerances to cold, salt, acid, or desiccation stress.* Frontiers in microbiology, 2017. **8**: p. 369.
49. Cao, M.D., et al., *Scaffolding and completing genome assemblies in real-time with nanopore sequencing.* Nature communications, 2017. **8**(1): p. 14515.
50. Phillippy, A.M., M.C. Schatz, and M. Pop, *Genome assembly forensics: finding the elusive mis-assembly.* Genome biology, 2008. **9**: p. 1-13.
51. Beatson, S.A. and M.J. Walker, *Tracking antibiotic resistance.* Science, 2014. **345**(6203): p. 1454-1455.
52. Bashir, A., et al., *A hybrid approach for the automated finishing of bacterial genomes.* Nature biotechnology, 2012. **30**(7): p. 701-707.
53. Bolger, A.M., M. Lohse, and B. Usadel, *Trimmomatic: a flexible trimmer for Illumina sequence data.* Bioinformatics, 2014. **30**(15): p. 2114-2120.
54. Bankevich, A., et al., *SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing.* Journal of computational biology, 2012. **19**(5): p. 455-477.
55. Souvorov, A., R. Agarwala, and D.J. Lipman, *SKESA: strategic k-mer extension for scrupulous assemblies.* Genome biology, 2018. **19**(1): p. 1-13.
56. Li, R., et al., *Efficient generation of complete sequences of MDR-encoding plasmids by rapid assembly of MinION barcoding sequencing data.* Gigascience, 2018. **7**(3): p. gix132.
57. Wick, R.R., et al., *Unicycler: resolving bacterial genome assemblies from short and long sequencing reads.* PLoS computational biology, 2017. **13**(6): p. e1005595.
58. Walker, B.J., et al., *Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement.* PloS one, 2014. **9**(11): p. e112963.
59. Davis, S., et al., *CFSAN SNP Pipeline: an automated method for constructing SNP matrices from next-generation sequence data.* PeerJ Computer Science, 2015. **1**: p. e20.

60. Langmead, B. and S.L. Salzberg, *Fast gapped-read alignment with Bowtie 2.* Nature methods, 2012. **9**(4): p. 357-359.
61. Koboldt, D.C., et al., *VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing.* Genome research, 2012. **22**(3): p. 568-576.
62. Chen, Y., et al., *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.* Applied and Environmental Microbiology, 2017. **83**(15).
63. Kumar, S., et al., *MEGA X: molecular evolutionary genetics analysis across computing platforms.* Molecular biology and evolution, 2018. **35**(6): p. 1547.
64. Arndt, D., et al., *PHASTER: a better, faster version of the PHAST phage search tool.* Nucleic acids research, 2016. **44**(W1): p. W16-W21.
65. Carver, T.J., et al., *ACT: the Artemis comparison tool.* Bioinformatics, 2005. **21**(16): p. 3422-3423.
66. Carattoli, A., et al., *In silico detection and typing of plasmids using PlasmidFinder and plasmid multilocus sequence typing.* Antimicrobial agents and chemotherapy, 2014. **58**(7): p. 3895-3903.
67. Darling, A.C., et al., *Mauve: multiple alignment of conserved genomic sequence with rearrangements.* Genome research, 2004. **14**(7): p. 1394-1403.
68. Bouckaert, R., et al., *BEAST 2: a software platform for Bayesian evolutionary analysis.* PLoS computational biology, 2014. **10**(4): p. e1003537.
69. Rambaut, A., et al., *Posterior summarization in Bayesian phylogenetics using Tracer 1.7.* Systematic biology, 2018. **67**(5): p. 901-904.
70. Orsi, R.H., et al., *Short-term genome evolution of Listeria monocytogenes in a non-controlled environment.* BMC genomics, 2008. **9**: p. 1-17.
71. Kuenne, C., et al., *Reassessment of the Listeria monocytogenes pan-genome reveals dynamic integration hotspots and mobile genetic elements as major components of the accessory genome.* BMC genomics, 2013. **14**(1): p. 1-19.
72. Marik, C.M., et al., *Growth and survival of Listeria monocytogenes on intact fruit and vegetable surfaces during postharvest handling: A systematic literature review.* Journal of food protection, 2020. **83**(1): p. 108-128.
73. CDC. *Multistate outbreak of listeriosis linked to whole cantaloupes from Jensen Farms, Colorado (final update)*. 2011; Available from: Available at: https://www.cdc.gov/listeria/outbreaks/cantaloupes-jensen-farms/index.html.
74. Chen, Y., et al., *Listeria monocytogenes in stone fruits linked to a multistate outbreak: enumeration of cells and whole-genome sequencing.* Applied and environmental microbiology, 2016. **82**(24): p. 7030-7040.
75. CDC. *Multistate outbreak of listeriosis linked to commercially produced, prepackaged caramel apples made from Bidart Bros. apples (final update)*. 2015; Available from: Available at: http://www.cdc.gov/listeria/outbreaks/caramel-apples12-14/.
76. FDA. *FY 2014 – 2016 Microbiological Sampling Assignment Summary Report: Whole Fresh Avocados*. 2018; Available from: Available at: https://www.fda.gov/food/sampling-protect-food-supply/microbiological-surveillance-sampling-fy14-16-whole-fresh-avocados.

77. Gurevich, A., et al., *QUAST: quality assessment tool for genome assemblies.* Bioinformatics, 2013. **29**(8): p. 1072-1075.
78. Ruppitsch, W., et al., *Defining and Evaluating a Core Genome Multilocus Sequence Typing Scheme for Whole-Genome Sequence-Based Typing of Listeria monocytogenes.* Journal of Clinical Microbiology, 2015. **53**(9): p. 2869-2876.
79. Gorski, L., et al., *Prevalence and Clonal Diversity of over 1,200 Listeria monocytogenes Isolates Collected from Public Access Waters near Produce Production Areas on the Central California Coast during 2011 to 2016.* Applied and Environmental Microbiology, 2022.
80. Allard, M.W., et al., *Whole genome sequencing uses for foodborne contamination and compliance: discovery of an emerging contamination event in an ice cream facility using whole genome sequencing.* Infection, Genetics and Evolution, 2019. **73**: p. 214-220.
81. Wang, Y., et al., *Genetic diversity of Salmonella and Listeria isolates from food facilities.* Journal of food protection, 2018. **81**(12): p. 2082-2089.
82. Ragon, M., et al., *A new perspective on Listeria monocytogenes evolution.* PLoS Pathog, 2008. **4**(9): p. e1000146.
83. Maury, M.M., et al., *Hypervirulent Listeria monocytogenes clones' adaption to mammalian gut accounts for their association with dairy products.* Nature communications, 2019. **10**(1): p. 1-13.
84. Piercey, M.J., P.A. Hingston, and L. Truelstrup Hansen, *Genes involved in Listeria monocytogenes biofilm formation at a simulated food processing plant temperature of 15 °C.* Int J Food Microbiol, 2016. **223**: p. 63-74.
85. Kuenne, C., et al., *Comparative analysis of plasmids in the genus Listeria.* PloS one, 2010. **5**(9): p. e12511.
86. Lee, S., et al., *The arsenic resistance-associated Listeria genomic island LGI2 exhibits sequence and integration site diversity and a propensity for three Listeria monocytogenes clones with enhanced virulence.* Applied and environmental microbiology, 2017. **83**(21).
87. Sauders, B.D., et al., *Molecular characterization of Listeria monocytogenes from natural and urban environments.* Journal of food protection, 2006. **69**(1): p. 93-105.
88. Den Bakker, H.C., et al., *Lineage specific recombination rates and microevolution in Listeria monocytogenes.* BMC evolutionary biology, 2008. **8**(1): p. 1-13.
89. Palacios-Gorba, C., et al., *Ruminant-associated Listeria monocytogenes isolates belong preferentially to dairy-associated hypervirulent clones: a longitudinal study in 19 farms.* Environmental Microbiology, 2021. **23**(12): p. 7617-7631.
90. Jacquet, C., et al., *A molecular marker for evaluating the pathogenic potential of foodborne Listeria monocytogenes.* The Journal of infectious diseases, 2004. **189**(11): p. 2094-2100.
91. Wang, J., et al., *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 pathogens and disease, 2015. **12**(2): p. 151-158.
92. FDA. *FY 2014 – 2016 Microbiological Sampling Assignment Summary Report: Raw Milk Cheese Aged 60 Days*. 2016; Available from: Available at: https://www.fda.gov/food/sampling-protect-food-supply/microbiological-surveillance-sampling-fy14-16-raw-milk-cheese-aged-60-days.
93. FDA. *FY 2014 – 2016 Microbiological Sampling Assignment Summary Report: Sprouts*. 2017; Available from: Available at: https://www.fda.gov/food/sampling-protect-food-supply/microbiological-surveillance-sampling-fy14-16-sprouts.
94. Control, C.f.D. and Prevention, *Multistate outbreak of listeriosis linked to Blue Bell Creameries products (final update)*. 2015.
95. Chen, Y., et al., *Assessing the genome level diversity of Listeria monocytogenes from contaminated ice cream and environmental samples linked to a listeriosis outbreak in the United States.* PLoS One, 2017. **12**(2): p. e0171389.
96. CDC. *Multistate Outbreak of Listeriosis Linked to Blue Bell Creameries Products (Final Update)*. 2015; Available from: Available at: https://www.cdc.gov/listeria/outbreaks/ice-cream-03-15/index.html.
97. FDA, *Inspection and Environmental Sampling of Ice Cream Production Facilities for Listeria monocytogenes and Salmonella FY 2016-17.* 2022.
98. Li, Z., et al., *Whole genome sequencing analyses of Listeria monocytogenes that persisted in a milkshake machine for a year and caused illnesses in Washington State.* BMC microbiology, 2017. **17**(1): p. 1-11.
99. Vielva, L., et al., *PLACNETw: a web-based tool for plasmid reconstruction from bacterial genomes.* Bioinformatics, 2017. **33**(23): p. 3796-3798.
100. Schmitz-Esser, S., J.M. Anast, and B.W. Cortes, *A large-scale sequencing-based survey of plasmids in Listeria monocytogenes reveals global dissemination of plasmids.* Frontiers in microbiology, 2021. **12**: p. 510.
101. Calley, B.B., et al., *A comparative content analysis of news stories and press releases during the 2015 blue bell ice cream recall.* Journal of applied communications, 2019. **103**(3).