CENTER for PRODUCE SAFETY

CPS 2017 RFP
FINAL PROJECT REPORT

Project Title
Listeria whole genome sequence data reference sets are needed to allow for improved persistence assessment and source tracking

Project Period
January 1, 2018 – December 31, 2019

Principal Investigator
Martin Wiedmann
Cornell University
Department of Food Science
341 Stocking Hall
Ithaca, NY 14853
T: 607-254-2838
E: mw16@cornell.edu

Objectives

1. *Develop a sampling plan for collection, across the US, of at least 1,500* soil samples focusing on non-agricultural and natural environments, followed by testing of samples for Listeria monocytogenes and Listeria spp.*
2. *Perform whole genome sequencing (WGS) of the L. monocytogenes and Listeria spp. isolates obtained through Obj. 1, and analyze data to assess associations between WGS sequence type and geographical origin.*
3. *Perform WGS of Listeria spp. isolated from throughout the produce chain (for example from irrigation water, packinghouses, processing facilities, and produce environments in retail stores); isolates will be obtained from pre-existing isolate collections, and through concurrent sampling efforts that are part of ongoing, funded studies.*
4. *Perform a comprehensive analysis of Listeria spp. and L. monocytogenes WGS data to provide information on the number of single nucleotide polymorphism (SNP) or allelic differences that provide an appropriate cut-off to identify isolates with a likely epidemiological link.*

(* Revised to 1,000 composite soil samples.)

Funding for this project provided by the Center for Produce Safety through:
Florida Department of Agriculture and Consumer Services (FDACS) SCBGP grant# 24842

FINAL REPORT

Abstract

Whole genome sequencing (WGS) of *Listeria monocytogenes* (*LM*) has been used for routine human foodborne disease surveillance in the US since 2013. Regulatory agencies also routinely use WGS to characterize *LM* isolates obtained from foods, food processing facilities, and food-associated environments. Despite considerable WGS work on human isolates, there are currently limited data on the distribution and diversity of *LM* and *Listeria* WGS-based subtypes in non-food associated environments. Interpretation of WGS data hence does not have the benefit of comparison data that could be used to assess the likelihood of closely related *LM* and *Listeria* spp. being isolated from different sources. In this project, we developed a sampling plan and collected more than 1,000 soil samples across the US, focusing on non-agricultural and natural environments, followed by testing of samples for *LM* and *Listeria* spp. and performing WGS on representative isolates. A total of 1,854 *sigB*-sequencing-confirmed *Listeria* isolates including 659 *LM* isolates were obtained from 311 soil samples (prevalence 31%), and 594 *Listeria* isolates including 177 *LM* isolates were selected as representative isolates for WGS. From ongoing and previous projects which study *Listeria* in food associated environments, we selected and performed WGS on 235 *Listeria* spp. isolates (56 *L. innocua,* 21 *L. marthii,* 101 *L. seeligeri, and* 57 *L. welshimeri*) isolated from irrigation water and processing facilities. Core genome multilocus sequence typing (cgMLST) analysis and high-quality single nucleotide polymorphism (hqSNP) analysis were performed to access the relatedness between *LM* from non-food associated environments and from clinical cases, and between *Listeria* spp. from soil in non-food associated environments and in food associated environments.

Results showed that the diversity of *Listeria* in soil was very high, with 14 known species and 6 potential new species identified. *L. monocytogenes* was the most prevalent species, followed by *L. welshimeri* and *L. seeligeri*. *L. seeligeri* was only found in the northern US. *L. booriae* was isolated for the first time from soil and ranked fourth in prevalence; it had a surprisingly large genome size (3.5-4.0 Mbp). Three lineages of *LM* were detected from soil: lineage I, II and III, with lineage III found to be the most prevalent. Overall, *Listeria* in soil across the contiguous US showed a distribution delineated by longitude and elevation and mainly driven by soil moisture and sodium content. *LM* and *Listeria* spp. exhibited different geographic distribution patterns and environment preference in terms of soil property, climate and land-use. A total of 6 *LM* from soil were found to be closely related to clinical isolates submitted to NCBI (<20 cgMLST mismatches). *Listeria* spp. from non-food associated environments and from food-associated environments were not very closely related except for *L. seeligeri*. A total of 8 *L. seeligeri* isolate pairs showed less than 20 hqSNP differences. Our study improved the understanding of the distribution and ecology of *LM* and *Listeria* spp. across the US and suggested that soil in non-food associated environments may not be a common source for *Listeria* pathogen infection in humans and *Listeria* contamination in the produce industry, except for *L. seeligeri*. WGS data show that *LM* isolates from natural environments may differ from human isolates by as few as 10 cgMLST alleles, which is further evidence that low levels of SNP differences between human and food or environmental isolates does not necessarily imply a causal link.

Background

Whole genome sequencing (WGS) of *Listeria monocytogenes* (*LM*) has been used for routine human foodborne disease surveillance in the US since 2013. Regulatory agencies also routinely use WGS to characterize *LM* isolates obtained from foods, food processing facilities, and food-associated environments to support surveillance as well as identification of outbreaks and outbreak sources. WGS data have also been used to identify persistence of *LM* and *Listeria* spp. strains in processing facilities. In several cases, the repeated isolation of identical *LM* or *Listeria* spp. subtypes (as determined by WGS or pulsed-field gel electrophoresis[PFGE]) has been interpreted as persistence and used as an indicator of unsanitary conditions that are associated with an increased risk of finished product contamination. It is noteworthy that this logic has been used in at least some recalls.

While WGS has significantly improved discriminatory power over PFGE, and therefore provides for improved subtyping, the interpretation of WGS data can be challenging, particularly if isolates differ by a small number of SNPs (single nucleotide polymorphisms). While *LM* and *Listeria* isolates may rapidly accumulate SNPs, data collected to date generally indicate that SNP differences arise, on average, at a rate of around 1 core genome SNP every 1 to 2 years (Moura et al., 2016). Hence, it is conceivable that *Listeria* isolates that differ by a small number of SNPs (for example <5 SNP differences) may share a recent common ancestor that occurred 5 to 10 years ago. Consequently, isolation of *LM* or *Listeria* with identical or nearly identical genomes (e.g., <5 SNP differences) from two different locations does not necessarily imply a common ancestor that is recent enough to support a causal link in a trace-back study or outbreak investigation. For example, our research group recently reported that *LM* with genome sequences that differed by <5 SNPs were isolated from retail stores with different ownership in three non-contiguous US states, suggesting a common source “upstream” in the system (such as a processing plant) (Stasiewicz et al., 2015). These findings clearly illustrate some of the potential challenges around interpretation of WGS data, as finding human cases with an *LM* isolate that matches an isolate from one of these stores could have been misinterpreted as suggesting this retail store as the outbreak source, even though the outbreak could have easily been caused by another retail store or a different source. Similarly, repeated isolation of this specific WGS in one retail facility could have been interpreted as “persistence” when it may have represented re-introduction, for example from a supplier.

As detailed above, determination of whether two or more isolates are “identical” or “closely related” will play an increasingly important role. In many cases the “cut-offs” used are empirical and based on experience; ultimate decisions in outbreak and trace-back investigations require support by both epidemiological and WGS evidence. Initial efforts are under way to better define cut-offs that can be used to predict whether isolates are closely enough related to likely be linked to the same source or outbreak. Currently, WGS data are typically analyzed using two different approaches to determine whether isolates are “identical” (defined as no detectable differences) or closely related (defined as difference below a certain threshold): (i) core genome (cg)MLST approach (routinely used by CDC) (Chen et al., 2016) and (ii) high-quality (hq) SNP approach (routinely used by US FDA) (Jackson et al., 2016a); this project analyzed the data collected using both approaches. For cgMLST, differences are measured as the number of “allelic mismatches”; for *LM* this is defined as the number of loci, among 1,748 loci, that do not match. Simply speaking, one allelic mismatch means that two isolates are identical for 1,747 genes, but differ at one gene. For hqSNP, differences are measured as the number of SNP differences (meaning the number of nucleotides that differ) between two isolates based on a total of about 2.7 million nucleotides (Jackson et al., 2016b). A recent publication that analyzed WGS data for 1,696 human isolates from across the world (Moura et al., 2016) provided an initial attempt to better define allelic mismatch cut-offs that provide for meaningful classification. As detailed by Moura et al., “most isolates sampled during investigations of single outbreaks

had seven or fewer allelic mismatches. […]). Second, taking into account the entire data set, a sharp discontinuity was observed, with few pairs of isolates having between seven and ten allelic mismatches, showing that isolates with no documented epidemiological link differed most generally by more than ten mismatches." While this provides some initial data on possible cut-offs to define closely related isolates, as well as potential approaches that can be used to define cut-offs, further work, including on non-human isolates, is essential to facilitate judicious and improved use of these tools, particularly when characterizing produce associated isolates.

Despite considerable WGS work on human isolates, there are currently very limited data on the distribution and diversity of *LM* and *Listeria* WGS-based subtypes in non-food associated environments. Interpretation of WGS data obtained by regulatory and public health agencies, as well as industry, hence does not have the benefit of comparison data that could be used to assess the likelihood of closely related *LM* and *Listeria* spp. being isolated from different sources. Comprehensive WGS data for *LM* and *Listeria* spp. isolated from different geographic locations across the US are needed to assess whether WGS-based subtypes are associated with specific regions. Without these data an assumption that specific WGS-based subtypes may be specific to certain regions or a given field or facility may be incorrectly used in trace-back investigations for both produce contamination events and foodborne disease outbreaks. For example, repeat isolation of certain *Salmonella* subtypes in specific regions has been used to identify a given region (e.g., Delmarva) as a likely source of a given produce related outbreak. Consequently, there is a need for an improved understanding of the distribution and ecology of *LM* and *Listeria* spp. whole genome sequence-based subtypes across the US to optimize the use of WGS for source tracking and assessment of *LM* and *Listeria* spp. persistence.

To fulfill these needs, the project team collected *LM* and *Listeria* spp. isolates from throughout the US and from non-produce related sources to create a reference WGS data set. The development of a reference WGS data set for isolates from non-produce related environments is essential to allow for improved interpretation of WGS data for isolates obtained from produce related environments. Such a reference WGS data set was used to determine whether and how frequently isolates from produce associated environments may match isolates from other, highly likely unrelated, sources.

Research Methods and Results

Sampling and soil processing. We designed a sampling map that covers the contiguous US, with grids measuring 5° by 5° fitted to the map (Figure 1). This design yielded an initial set of 60 grids (i.e., 5 grids cover the N to S direction and 12 grids cover the E to W direction), but only those for which at least 1/3 of the grid is occupied by US land were considered as sampling grids. Since Maine was separated into two grids, these two grids were combined and designated as a single sampling grid. By following this approach, a total of 40 sampling grids were confirmed and assigned numbers 1 to 40 (Figure 1). In the sampling plan, each grid had 5 sampling areas, with each pair being >20 km apart. Sampling areas were defined as rural natural environments with minimal human disturbance, such as national/state parks, forests, preserves, wildlife management areas, nature conservancy, and natural areas in proximity to rivers, lakes, and natural ponds. Each sampling area had 5 sampling sites, with each pair being >0.2 km apart. Each sampling site had 3 sampling points randomly selected by collectors. These 3 sampling points needed to be about 6 m (20 feet) apart and have a similar eco-environment (e.g., soil type, plant cover). Topsoil (0–8 inches) was collected at each of the 3 sampling points and pooled for one composite sample per sampling site. The sampling units are summarized in Table 1.

Sampling took place from April to November 2018 following the sampling plan above. Most of the samples were collected on days with average daytime temperature of 15°C, in order

to increase the likelihood of detecting *Listeria* and reduce the influence of daily weather on the distribution of *Listeria*. Sample collectors were recruited from among Cornell alumni as well as collaborators and colleagues across the US, forming a sampling team with more than 80 people. A sampling kit including sampling tools and a pre-tested standard sampling protocol, which contained the information of sampling tools, detailed sampling methods, shipping procedures, field datasheet, and sampling map with suggested GPS coordinates created by ArcGIS version 10.2 for each sampling site, was distributed to each collector before sampling. Soils were collected with sterile scoops by collectors wearing sanitized gloves at or close to the suggested GPS coordinates for each sampling site based on the accessibility, and were placed in a well-sealed Whirl-Pak bag. Actual GPS coordinates, sampling date and time, habitat types (e.g., forest, savanna, shrubland, grassland), and other notes for each sampling site were recorded by collectors on the field datasheet. The sampling kit including soil samples on ice, field datasheet, and sampling materials were shipped back to the Food Safety Lab at Cornell University. Soils were processed immediately upon arrival.

After briefly homogenizing the soil, basic observations (presence of roots, vegetation, or insects, the total weight of soil, soil color, soil texture) for each soil sample were recorded. A total of 25 g of soil from each sample was transferred to a sterile pre-labeled filter Whirl-Pak bag prepared for *Listeria* enrichment and isolation. A total of 10 g of soil from each sample was tested for water content following methods described in Black (1965). The remaining soil from each sample was stored at 4°C for other soil property measurement.

In the end, a total of 1,045 soil samples were collected from 209 sampling areas representing non-agricultural and natural environments across the US. After removing samples that did not meet the distance criteria (i.e., >20 km distance for sampling areas and >0.2 km for sampling sites), 1,004 samples were retained for the data analyses.

***Listeria* enrichment, isolation, and *sigB* sequencing.** *Listeria* enrichment and isolation were conducted as described previously (Weller et al., 2015). Briefly, for each soil sample, 225 mL of buffered *Listeria* enrichment broth (BLEB) was added into each sterile filter Whirl-Pak bag containing 25 g of soil. After 4 h, 900 uL of *Listeria* selective enrichment supplement (LSES) was added to each sample enrichment bag. At 24 and 48 h, 50 µl of each sample enrichment was streaked onto *Listeria monocytogenes* plating medium agar (LMPM) and Modified Oxford agar (MOX) plates. Following incubation of LMPM plates at 35°C and MOX plates at 30°C for 48h, up to 8 presumptive *L. monocytogenes* and/or *L. ivanovii* (blue colonies) and other presumptive *Listeria* species (white colonies) on LMPM, and *Listeria* (black colonies) on MOX were sub-streaked onto brain-heart infusion (BHI) agar plates. BHI plates were incubated at 37°C for 24 h (note: for some slow growers, incubation time was up to 72 h). Presumptive *Listeria* colonies selected from BHI plates were confirmed by *sigB* sequencing as detailed below. All *Listeria* isolates have been preserved at −80°C, and metadata has been deposited to Food Microbe Tracker (FMT).

Polymerase chain reaction (PCR) amplification of *sigB* for presumptive *Listeria* isolates was performed using the regular *sigB* primers – Lm sigB15 (5'-AATATATTAATGAAAAGCAGGTGGAG-3') and Lm sigB16 (5'-ATAAATTATTTGATTCAACTGCCTT-3') – according to the protocol provided in Liao et al. (2017) with minor modification on the concentration of primers (12.5µM). If any isolates showed abnormal bands (lower molecular weight bands) or no amplification, *sigB* degenerate PCR was performed on these isolates using the degenerate primers – sigBdegF (5'-TCVMAAGGYAAAWSYTTYCAYGARGA-3') and sigBdegR (5'-GASACRTGCATTTGWGATATAYCGAG-3') – based on protocol provided on Food Safety Wiki. *sigB* sequencing was performed at the Genomics Facility of the Cornell University Institute of Biotechnology. The *sigB* allelic type (AT) of each *Listeria* isolate was identified by comparing *sigB* sequences to an internal reference database using Basic Local Alignment Search Tool

(BLAST). *sigB*-confirmed *Listeria* isolates having distinct *sigB* AT within each sample were selected for WGS. If multiple isolates had the same AT within each sample, one isolate was randomly selected as the representative isolate.

A total of 1,854 *sigB*-sequencing-confirmed *Listeria* isolates, including 659 *LM* isolates, were obtained from 311 soil samples (prevalence 31%). A total of 594 *Listeria* isolates, including 177 *LM* isolates, were selected as representative isolates from non-food associated environments for WGS.

***Listeria* spp. from food associated environments.** Specialty crop processing facilities and agricultural water were selected to represent food associated environments in this project. *Listeria* spp. isolated from specialty crop processing facilities originated from an ongoing project "*Development of Listeria control strategies for specialty crop processing facilities*," while *Listeria* spp. isolated from agricultural water originated from previous projects published in Weller et al. (2019, 2020). In brief, isolates from processing facilities represented zones 2–4 samples from 13 produce packinghouses and 3 fresh-cut produce operations in the United States, collected between August 2017 and October 2018. Isolates from agricultural water were collected using a set of Moore swabs and a set of grab samples in Arizona between February and December 2017 and New York between May and September 2017 and in 2018; these isolates included the species *L. marthii*, *L. seeligeri*, *L. innocua*, and *L. welshimeri*. WGS was performed on a total of 235 *Listeria* spp. isolates from food associated environments, and included 56 *L. innocua,* 21 *L. marthii,* 101 *L. seeligeri, and* 57 *L. welshimeri*.

WGS, genome assembly, hqSNP, and cgMLST. WGS was performed on representative *Listeria* isolated from soil and food associated environments. The total DNA of isolates was extracted using the QIAamp DNA MiniKit (Qiagen, Valencia, CA, USA) according to the manufacturer's protocol. DNA quality was assessed using Nanodrop, and DNA concentration was measured using Qubit. DNA products with A260/280 of ~1.80 and A260/230 of ~2.0 were considered good quality and constructed for Nextera XT libraries (Illumina, Inc., San Diego, CA, USA). Sequencing was performed on multiple platforms, including Illumina MiSeq with 250 bp paired end reads at the Cornell Vet School, Illumina HiSeq and NextSeq with both 150 bp paired end reads at the Genomics Facility of the Cornell University Institute of Biotechnology, and Illumina NextSeq with 150 bp paired end reads at New York State Department of Health.

Genome assembly was performed following methods as described in Kovac et al.(2017). After adapters and low-quality bases were trimmed by Trimmomatic 0.39 (Bolger et al., 2014), paired-end reads were assembled de novo using SPAdes 3.13.1 with kmer sizes of 21, 33, 55, 77 for Illumina HiSeq and NextSeq, and 33, 55, 77, 99,127 for Illumina MiSeq and a coverage cutoff of 2.0 (Bankevich et al., 2012). Contigs >500 bp were removed. All assemblies had passed the quality control (number of contigs >300, reasonable genome size, N50 >50000, average coverage >30x, consistent WGS-extracted *sigB* AT and PCR-based *sigB* AT) and contamination screening using Kraken (Wood and Salzberg, 2014). Whole-genome pairwise average nucleotide and *sigB* AT were used to classify *Listeria* species.

To select closely related *Listeria* spp. from non-food and food associated environments for hqSNP analysis, core SNPs within *L. marthii*, *L. seeligeri*, *L. innocua*, and *L. welshimeri* were first identified by kSNP3, using a kmer size of 19, determined by KChooser (Gardner et al., 2015), and used to construct a maximum likelihood (ML) phylogenetic of each *Listeria* spp. using the GTRGAMMAX substitution model with ascertainment bias correction in RaxML version 8.2.12 (Stamatakis, 2014). ML bootstrap values were calculated based on 1,000 bootstrap repetitions. *Listeria* spp. isolate pairs with a core SNP difference <10 were selected for hqSNP analysis; hqSNP analysis was performed on closely related *Listeria* spp. pairs using CFSAN SNP pipeline v. 1.0.0 (Davis et al., 2015) with default parameters. The isolate which had higher assembly quality in the pair was selected as the reference for read mapping. Closely

related *LM* from non-food associated environments and clinical cases were identified when they were placed in clusters by single-linkage clusters with a maximum 50 SNP distance in NCBI Pathogens Isolates Browser (https://www.ncbi.nlm.nih.gov/pathogens/isolates#/search/). Whole genome assemblies of the most closely related clinical *LM* to each soil *LM* in the SNP cluster were downloaded from NCBI. cgMLST analysis was further performed on these closely related soil *LM* and clinical *LM* using an in-house pipeline to assign alleles based on the cgMLST database described by Moura et al. (2016) and available at https://bigsdb.pasteur.fr/listeria.

Soil property, climate, land-use and spatial data. Soil properties of samples positive for *Listeria* and a matching number of samples negative for *Listeria* were measured at the Cornell Nutrient Analysis Lab. The negative samples were determined using a customized random algorithm. Soil properties included total carbon (TC), total nitrogen (TN), organic matter, pH, aluminum (Al), calcium (Ca), copper (Cu), iron (Fe), potassium (K), magnesium (Mg), manganese (Mn), molybdenum (Mo), sodium (Na), phosphorus (P), sulfur (S), zinc (Zn). Climate data, including average precipitation, maximum temperature, minimum temperature, and wind speed from 1915–2011 were obtained from the National Oceanic and Atmospheric Administration (https://www.esrl.noaa.gov/psd/data/gridded/data.livneh.html) for each sampling site. To characterize the land cover surrounding each sampling site, we used inverse-distance weighting (IDW) as proposed by King et al. (2005) and previously implemented by Weller et al. (2019). Briefly, the IDW is based on the idea that land cover in areas closer to the sampling site will have a greater impact than that in areas farther away. The IDW proportion of the total area under each land cover class was calculated. Land cover classes included open water, developed open space (<20% impervious cover), developed (>20% impervious cover), barren, forest, shrubland, grassland, cropland, pasture, and wetland. Inverse distance weights were calculated using the following distance intervals: 0–50 m, 50–100m, 100–500m, 500–,000m, 1,000–5,000m, and 5,000–10,000m. Elevation was obtained from United States Geological Survey (https://viewer.nationalmap.gov/theme/elevation/#/%23bottom**)** for each sampling site.

Diversity, distribution, and ecology of *Listeria* in soil. The diversity of *Listeria* in soil was very high, with 20 species identified (all 6 known *sensu stricto* species, 8 known *sensu lato* species, 5 new *sensu stricto* species, and 1 new *sensu lato* species) (Figure 2). *L. monocytogenes* was the most prevalent species, with a prevalence of 11.75%, followed by *L. welshimeri and L. seeligeri*. *L. booriae* was discovered for the first time from soil and ranked fourth in prevalence. The genome size of *L. booriae* was surprisingly large (3.5–4.0 Mbp). Three lineages of *LM* were detected from soil: lineage I, II and III. *LM* lineage III, which was considered to be not common in the environment, turned out to be the most prevalent *LM* lineage, and occupied 77 sites (Figure 3).

Overall, *Listeria* in soil across the contiguous US showed a distribution significantly delineated by longitude and elevation (Figure 4). *Listeria* was much more prevalent in the east than in the west, and in regions with higher elevation (Figure 4b). Several noticeable hotspots of *Listeria*, such as areas around Ohio, Indiana, Alabama, and Iowa, were identified (Figure 4). *LM* and *Listeria* spp. exhibited different distribution patterns. *LM* distribution was enriched in the eastern US (Figure 5a), while the habitats of *L. welshimeri* ranged from the eastern to the central US (Figure 5b). *L. seeligeri* was only found in the northern US (Figure 5c), suggesting a preference for cold temperatures. *L. booriae* was widely distributed in the environment, especially in the south (Figure 5d); *L. innocua* was mainly distributed in the central US (Figure 5e).

A number of abiotic variables that were significantly different between samples positive and negative for *Listeria* were identified using the Mann-Whitney test (Figure 6; $p < 0.05$). Abiotic variables with $p < 0.001$ after false discovery rate (FDR) were defined as key abiotic factors. These variables included 10 soil variables (moisture, organic matter, total carbon, total nitrogen, copper, magnesium, manganese, molybdenum, sodium, and zinc) (Figure 6a), 1

climate variable (wind speed) (Figure 6b), and 8 land-use variables (pasture, grassland, shrubland, cropland, wetland, open water, and developed land) (Figure 6c). These key abiotic factors were included in Random Forest analysis to assess the importance of each variable in predicting the presence of *Listeria*. Results showed that sodium, moisture, and molybdenum were the top three important variables (Figure 7). Sites with higher moisture, lower sodium, and higher molybdenum may be more suitable for *Listeria* to grow (Figure 6a).

The ecological niche in terms of soil property, climate, and land-use varied by *LM* lineages and *Listeria* spp. Analysis of variance (ANOVA) test showed that the majority of soil variables, climate variables, and land-use variables were significantly different among *LM* lineages and *Listeria* spp. Permutational multivariate analysis of variance (PERMANOVA) *post hoc* was performed on significant variables to test whether the centroids of *LM* lineages and *Listeria* spp. are equivalent between each group pair based on significant variables of soil, climate and land-use. Results showed that most *LM* lineages and *Listeria* spp. were significantly different from each other in ecological niche, especially for climate and land-use (Figure 8).

Relatedness of *LM* between non-food associated environment and clinical cases. Based on cgMLST, *LM* isolates in non-food associated environments were not closely related. More than 97% of *LM* had more than 1,000 cgMLST mismatches with each other within lineage (Figure 9). A small number of closely related *LM* isolates were identified and were found in geographically close sampling areas (Figure 9). By comparing to the clinical *LM* submitted to NCBI, 6 soil *LM* (4 lineage I and 2 lineage II isolates) identified were placed in clusters by single-linkage clusters with a maximum 50 SNP differences in the Pathogens Isolates Browser, suggesting close relatedness. Results of cgMLST on these closely related soil and clinical *LM* isolates showed that cgMLST mismatches ranged between 10 to 17, a few mismatches more than the threshold (7.4) of potential epidemiological link proposed by Moura et al. (2018). These soil *LM* isolates were sampled between June to October 2018, and 5 out of 6 were found in the eastern US. Metadata such as isolation location and time were missing for most of the clinical *LM* (Table 2).

Relatedness of *Listeria* spp. between non-food and food associated environments. Four *Listeria* spp. – *L. marthii, L. innocua, L. welshimeri, and L. seeligeri* – were included in the relatedness comparison between non-food and food associated environments. Maximum likelihood phylogenetic trees of these *Listeria* spp. showed that *L. innocua, L. marthii,* and *L. welshimeri* isolates were clustering by sources in general (Figure 10a, 10b, 10c), meaning, for example, that isolates from natural environments clustered separately from isolates from processing plants and packinghouses. Only *L. seeligeri* had multiple clusters with a mixture of isolates from soil, agricultural water and processing plants (Figure 10d). Based on the distribution of core SNP differences of isolates from different sources, the majority of *L. marthii* from soil and agricultural water differed by more than 9,000 core SNPs; all *L. marthii* from soil and processing facilities had more than 9,500 core SNP differences (Figure 11a). The majority of *L. innocua* from soil and agricultural water differed by more than 4,500 core SNPs from soil and processing facility isolates (Figure 11b). Nearly all *L. welshimeri* from soil and agricultural water and from soil and processing facility had more than 2,000 core SNP differences (Figure 11c). However, a few *L. seeligeri* from different sources exhibited low core SNP differences (Figure 11d). For example, soil isolate L7-1175 from New York state and isolate W9-1176 from agricultural water as well as isolate S11-0076 from a processing facility differed by 1 core SNP.

Listeria spp. from non-food and food associated environments which had a core SNP difference fewer than 10 were selected for hqSNP analysis. A total of 12 such closely related isolate pairs within *L. seeligeri* were identified. The hqSNP differences between non-food and food associated environments ranged from 7 to 40 for these closely related *L. seeligeri* isolates (Table 3), suggesting there is higher potential for *L. seeligeri* transmission between non-food and food associated environments than for other *Listeria* spp.

Outcomes and Accomplishments

All the objectives of the project have been fully achieved. We designed a first ever large-scale sampling plan for *Listeria* in non-food associated environments across the US. We collected more than 1,000 soil samples and yielded a collection of 659 *LM* and 1,195 *Listeria* spp. isolates. Our data have largely increased the diversity of *Listeria* found in soil, including new *Listeria* species and underrepresented species such as *L. booriae*. These strains have been banked and are available for future research. Importantly, we generated WGS data for 177 *LM* isolates and 417 *Listeria* spp. These data allowed for (i) unambiguous identification of strains; (ii) classification of strains based on their genetic relationships; and (iii) an improved understanding of the genetic diversity and geographic structure of *LM* and *Listeria*, including initial data on the likelihood of finding identical or closely related *LM* or *Listeria* spp. in different locations. Understanding the relationships between WGS types and geographic origin will help industry interpret WGS data used for source tracking and outbreak investigations, and will provide an initial database of *Listeria* WGS at a nationwide scale, which can also be used to provide information on presence of specific WGS types in different locations and regions.

The project also yielded data on baseline frequency of *LM* and *Listeria* spp. in soil samples across the US. We provided initial data on environmental factors that may affect the likelihood of *LM* and *Listeria* isolation across the US. These data are valuable to growers as they provide baseline data on *Listeria* frequency, which can be used to support that *Listeria* can be expected at a certain rate in fields, and will also help growers identify environmental factors that may increase the likelihood of *Listeria* isolation, which can be used for targeted implementation of control strategies.

In addition, the project yielded an initial data set on the relative WGS diversity of *Listeria* spp. isolates from food associated environments. These data provided further information on the likelihood of isolating closely related *LM* or *Listeria* spp. from different sources and/or geographically distinct areas. In addition, our data allowed us to initially assess the persistence, frequency and spatial pattern of *Listeria* spp. in different food associated environments, as supported by WGS data. These data provided initial information on the number of SNP differences that can be observed if specific strains persist over time, which can be used to further and more accurately define the substitution rates for the *Listeria* genomes, which has been estimated to be 1 substitution (meaning 1 SNP) every 2.5 years based on data for isolates from human outbreaks and cases.

Lastly, our comprehensive WGS data analyses on comparison between soil *LM* and clinical *LM*, and between *Listeria* spp. from soil and food associated environments allowed for a better definition of SNP or allelic difference cut-offs that suggest that isolates share a common ancestor that is recent enough to support an epidemiological relationship. While these data considerably improved interpretation of WGS data for produce associated isolates, even with these data the WGS data still needs to be interpreted along with epidemiological data, particularly when establishing the cause of an outbreak or a contamination event.

Summary of Findings and Recommendations

Key findings

- The diversity of *Listeria* in soil was very high; 20 species identified including all 6 known *sensu stricto* species, 8 known *sensu lato* species, 5 new *sensu stricto* species, and 1 new *sensu lato* species.
- *L. monocytogenes* was the most prevalent species with a prevalence of 11.75%, followed by *L. welshimeri and L. seeligeri*. *L. booriae* was for the first time discovered from soil and ranked fourth in prevalence. Three lineages of *LM* were detected from soil, lineage I, II and III, with lineage III to be the most prevalent *LM* lineage.
- *Listeria* in soil across the contiguous US showed a distribution significantly delineated by longitude and elevation; this distribution was driven by a combination of soil, climate, and land-use variables; sodium, moisture, and molybdenum were identified as the top three important drivers.
- The ecological niche represented by soil property, climate and land-use varied by *LM* lineages and *Listeria* spp.
- *LM* in soil were very genetically diverse; closely related *LM* were more likely to be found at close locations.
- A few *LM* lineage I and II isolates were closely related to clinical isolates. They may be epidemiologically linked based on the cgMLST results.
- *Listeria* spp. in non-food associated environments and food associated environments were not very closely related except for *L. seeligeri* based on core SNP and hqSNP. Soil in non-food associated environments may not be a common source for *Listeria* spp. contamination in the produce industry except for *L. seeligeri*.

Recommendations

- While *Listeria monocytogenes* is on average found frequently in soil from natural environments, hotspots of soil *Listeria* were identified in areas around Ohio, Indiana, Alabama, and Iowa.
- Large-scale data collection and analysis efforts can identify areas with increased risk of *LM* presence and predictive efforts that account for key abiotic factors such as sodium, moisture and molybdenum may allow for improved approaches to reduce introduction of *Listeria* into food associated environments.
- Different strategies may need to be adopted when controlling specific *LM* lineage and specific *Listeria* spp. as our results showed that they preferred different ecological niches.
- While prevalence of *LM* lineage I and II was not high in soil, they have risk in transmitting to human and cause infection.
- There is appears to be higher risk of introducing *L. seeligeri* closely related to isolates from natural environments into produce associated environments
- Case studies derived from this research indicate that closely related isolates (based on WGS data) can be found in natural environments and produce associated environments (for *L. seeligeri*) and in natural environments and human clinical cases (for *L. monocytogenes*).

Literature Cited

Bankevich, A., S. Nurk, D. Antipov, A. A. Gurevich, M. Dvorkin, A. S. Kulikov, et al. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J. Comput. Biol. 19:455-77.

Black CA. Methods of Soil Analysis: Part I Physical and mineralogical properties. American Society of Agronomy, Madison, Wisconsin, USA. 1965.

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

Chen, Y., N. Gonzalez-Escalona, T. S. Hammack, M. Allard, E. A. Strain, and E. W. Brown. 2016. Core genome multilocus sequence typing for the identification of globally distributed clonal groups and differentiation of outbreak strains of *Listeria monocytogenes*. Appl. Environ. Micro. 82: 6258-6272.

Davis, S., J. B. Pettengill, Y. Luo, J. Payne, A. Shpuntoff, H. Rand, and E. Strain. 2015. CFSAN SNP Pipeline: an automated method for constructing SNP matrices from next-generation sequence data. PeerJ Computer Science, 1, p.e20.

Gardner, S. N., T. Slezak, and B. G. Hall. 2015. kSNP3. 0: SNP detection and phylogenetic analysis of genomes without genome alignment or reference genome. Bioinformatics 17: 2877-2878.

Jackson, B. R., C. Tarr, E. Strain, K.A. Jackson, A. Conrad, H. Carleton, L.S. Katz, S. Stroika, L.H. Gould, R.K. Mody, and B.J. Silk. 2016a. Implementation of nationwide real-time whole-genome sequencing to enhance listeriosis outbreak detection and investigation. Clin. Infect. Dis. 63: 380-386.

Jackson, K. A., S. Stroika, L. S. Katz, J. Beal, E. Brandt, C. Nadon, A. Reimer, B. Major, A. Conrad, C. Tarr, and B. R. Jackson. 2016b. Use of whole genome sequencing and patient interviews to link a case of sporadic listeriosis to consumption of prepackaged lettuce. J. Food Prot. 79: 806-809.

King, R. S., M. E. Baker, D. F. Whigham, D. E. Weller, T. E. Jordan, P. F. Kazyak, and M. K. Hurd. 2005. Spatial considerations for linking watershed land cover to ecological indicators in streams. Ecol. Appl. 15:137–153.

Kovac, J., K. J. Cummings, L. D. Rodriguez-Rivera, L. M. Carroll, A. Thachil, and M. Wiedmann. 2017. Temporal genomic phylogeny reconstruction indicates a geospatial transmission path of *Salmonella* Cerro in the United States and a clade-specific loss of hydrogen sulfide production. Front. Microbiol. 8:737.

Liao, J., M. Wiedmann, and J. Kovac. Genetic stability and evolution of the *sigB* allele, used for *Listeria sensu strictos*ubtyping and phylogenetic inference. 2017. Appl. Environ. Microbiol. 83:e00306-17.

Moura, A., A. Criscuolo, H. Pouseele, M. M. Maury, A. Leclercq, C. Tarr, J. T. Björkman, T. Dallman, A. Reimer, V. Enouf, and E. Larsonneur. 2016. Whole genome-based population biology and epidemiological surveillance of Listeria monocytogenes. Nat. Microbiol. 2:16185.

Stamatakis, A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313.

Stasiewicz, M. J., H. Oliver, M. Wiedmann, and H. den Bakker. 2015. Whole genome sequencing allows for improved identification of persistent Listeria monocytogenes in food associated environments. Appl. Environ. Microbiol. 81: 6024-6037.

Weller, D. L., N. Brassill, C. M. Rock, R. Ivanek, E. Mudrak, E. Ganda, S. Roof, and M. Wiedmann. 2020. Complex interactions between weather, and microbial and physiochemical water quality impact the likelihood of detecting foodborne pathogens in agricultural water used for produce production. Front. Microbiol. 11:134.

Weller, D. L., A. Belias, H. Green, M. Wiedmann, and S. Roof. 2019. Landscape, water quality, and weather factors associated with an increased likelihood of foodborne pathogen contamination of New York streams used to source water for produce production. Front. Sustain. Food Syst. Frontiers 3:124.

Weller, D., M. Wiedmann, and L. K. Strawn. 2015. Spatial and temporal factors associated with an increased prevalence of *Listeria monocytogenes* in spinach fields in New York State. Appl. Environ. Microbiol. 81:6059-6069.

Wood, D. E., and S. L. Salzberg. 2014. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol.15:R46.

APPENDICES

Publications (in preparation)
Liao, J., S. Guo, D. Weller, O. X. Cordero, M. Wiedmann. Ecological mechanisms underpinning the nationwide biogeographic pattern of edaphic *Listeria. (in preparation)*

Liao, J., S. Guo, O. H. Renato, S. Pollak, O. X. Cordero, M. Wiedmann. Explaining phylogeography of edaphic *Listeria* from a perspective of population genomics. *(in preparation)*

Presentations
Wiedmann, M. 2018. *Listeria* whole genome sequence data reference sets are needed to allow for improved persistence assessment and source tracking. 2018 CPS Research Symposium, June 19-20, Charlotte, NC.

Wiedmann, M. 2019. *Listeria* whole genome sequence data reference sets are needed to allow for improved persistence assessment and source tracking. 2019 CPS Research Symposium, June 18-19, Austin, TX.

Budget Summary
Total funds awarded to this project were $356,975. All funds awarded for this project were spent by the end of the budget term. The team had sufficient funds to fully implement this project.

Figures and Tables (see below)

Figures 1–11 and Tables 1–3

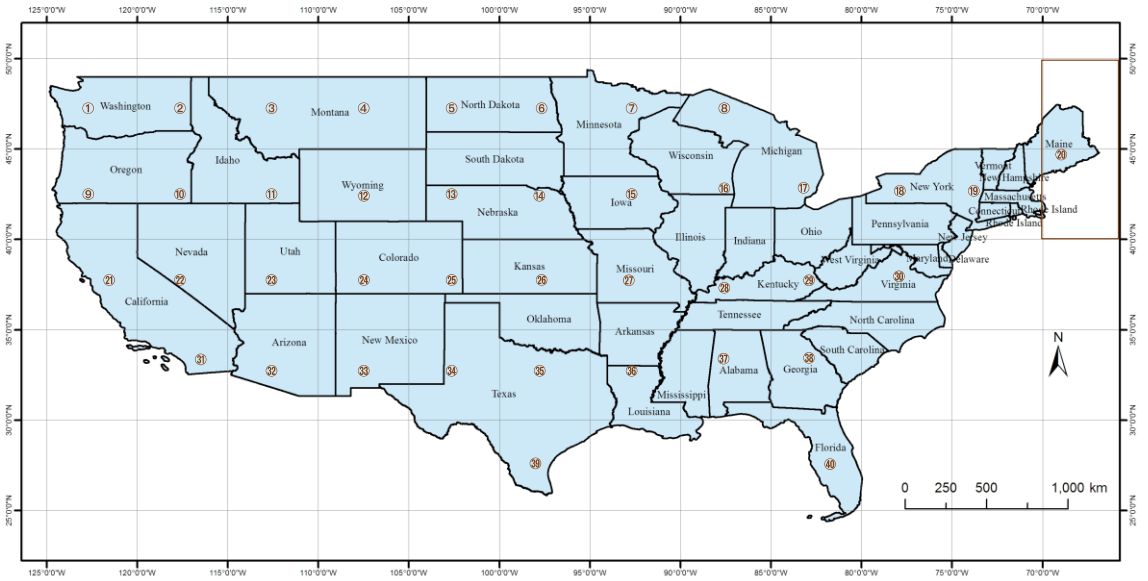

Figure 1. Sampling grids (n=40) used across continental US. The longitude and latitude coordinates of each intersection are provided. Dots in this map are the numbers assigned to each sampling grid, ranging from 1 to 40.

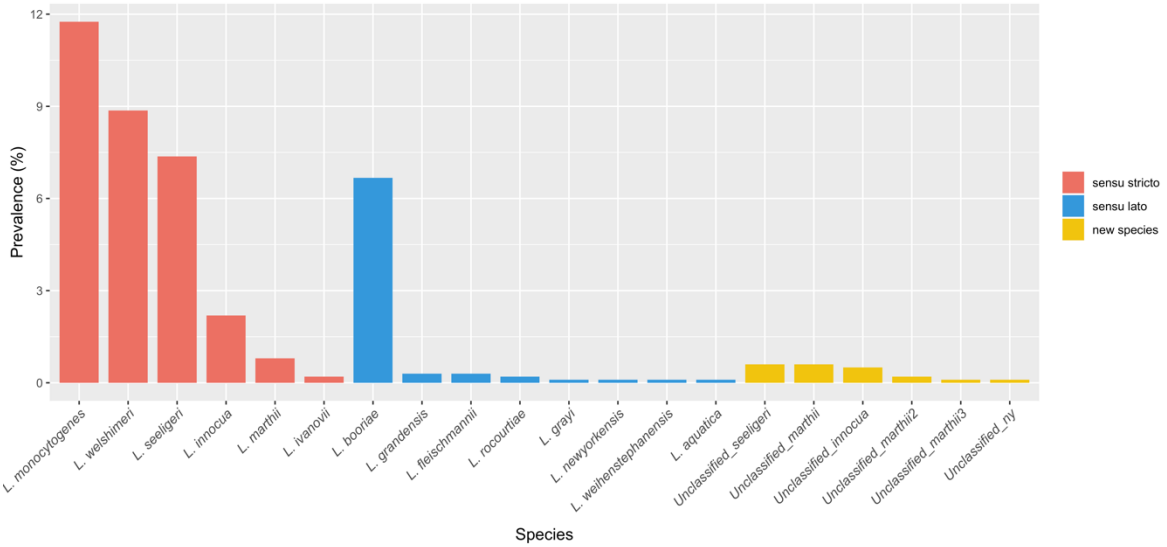

Figure 2. Prevalence of 20 species found in soil across the US. Prevalence is indicated by the number of sites that have certain species identified.

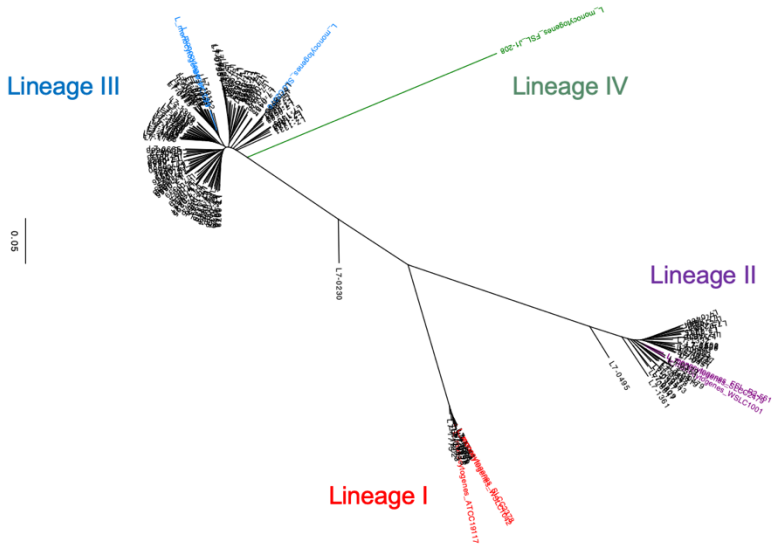

Figure 3. Unrooted phylogenetic tree of *LM* lineages based on core SNPs identified by kSNP3. Reference strains for each lineage were color coded.

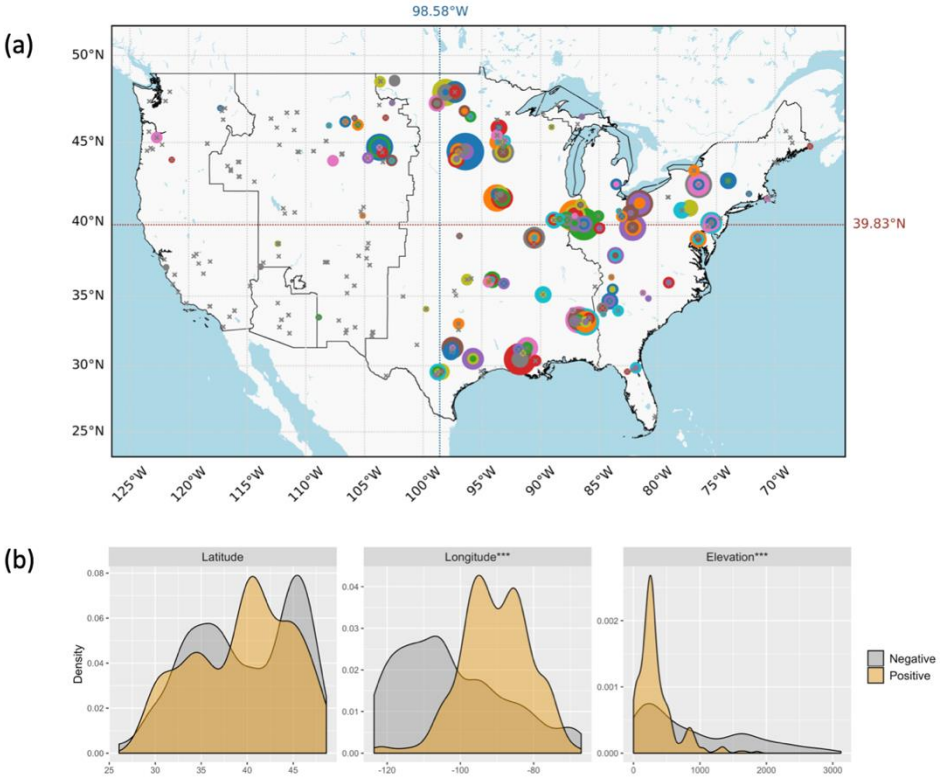

Figure 4. Distribution of *Listeria* across the US. (a) Map of sampling sites positive for *Listeria* (indicated by circles) and negative for *Listeria* (indicated by grey crosses). The circle color is random; circle size is in proportion to number of unique *Listeria* subtypes (*sigB* ATs). The red dashed line and blue dashed line are showing the latitude and longitude of the geographic center of the contiguous US. (b) Density of samples positive and negative for *Listeria* along latitude, longitude, and elevation. Spatial variables significantly different between positive samples and negative samples ($p < 0.001$ after FDR correction) using Benjamini and Hochberg by Mann-Whitney test are marked as "***".

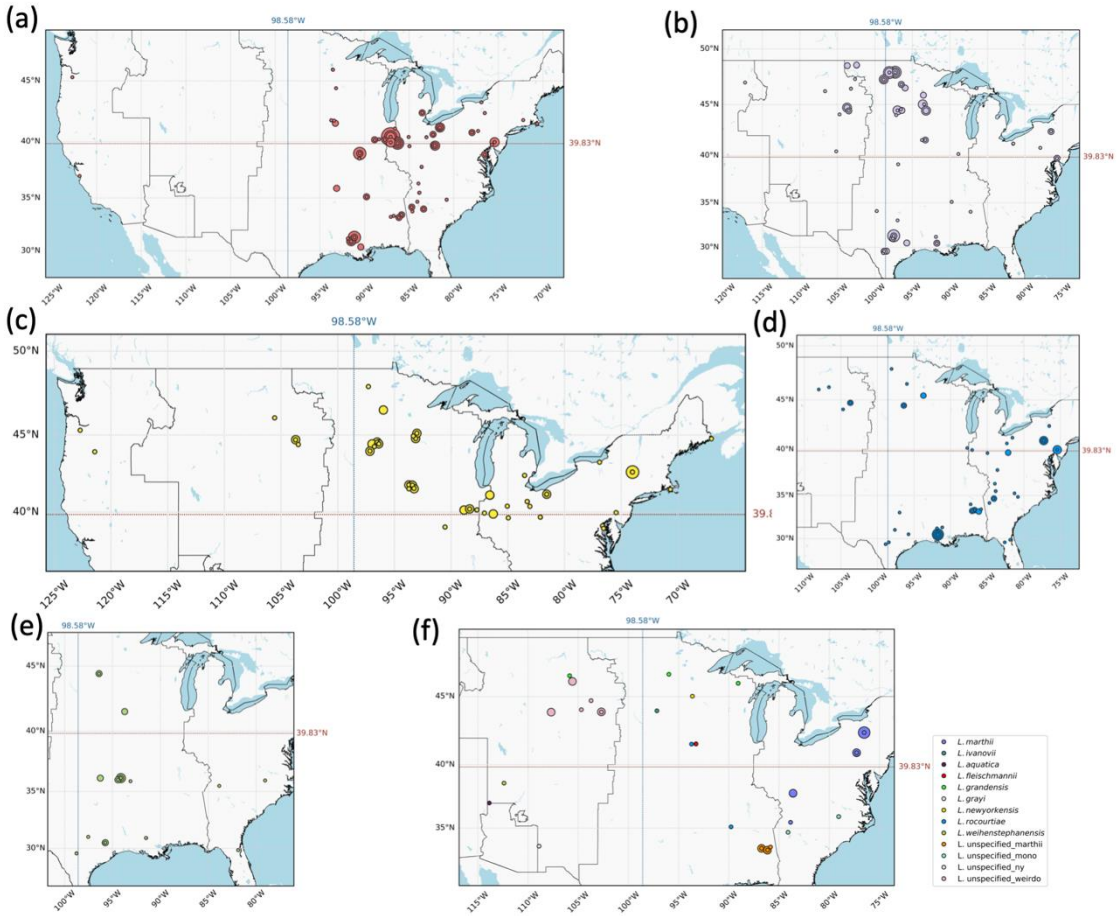

Figure 5. Distribution of (a) *LM*, (b) *L. welshimeri*, (c) *L. seeligeri*, (d) *L. booriae*, (e) *L. innocua*, and (f) other *Listeria* spp. across the US.

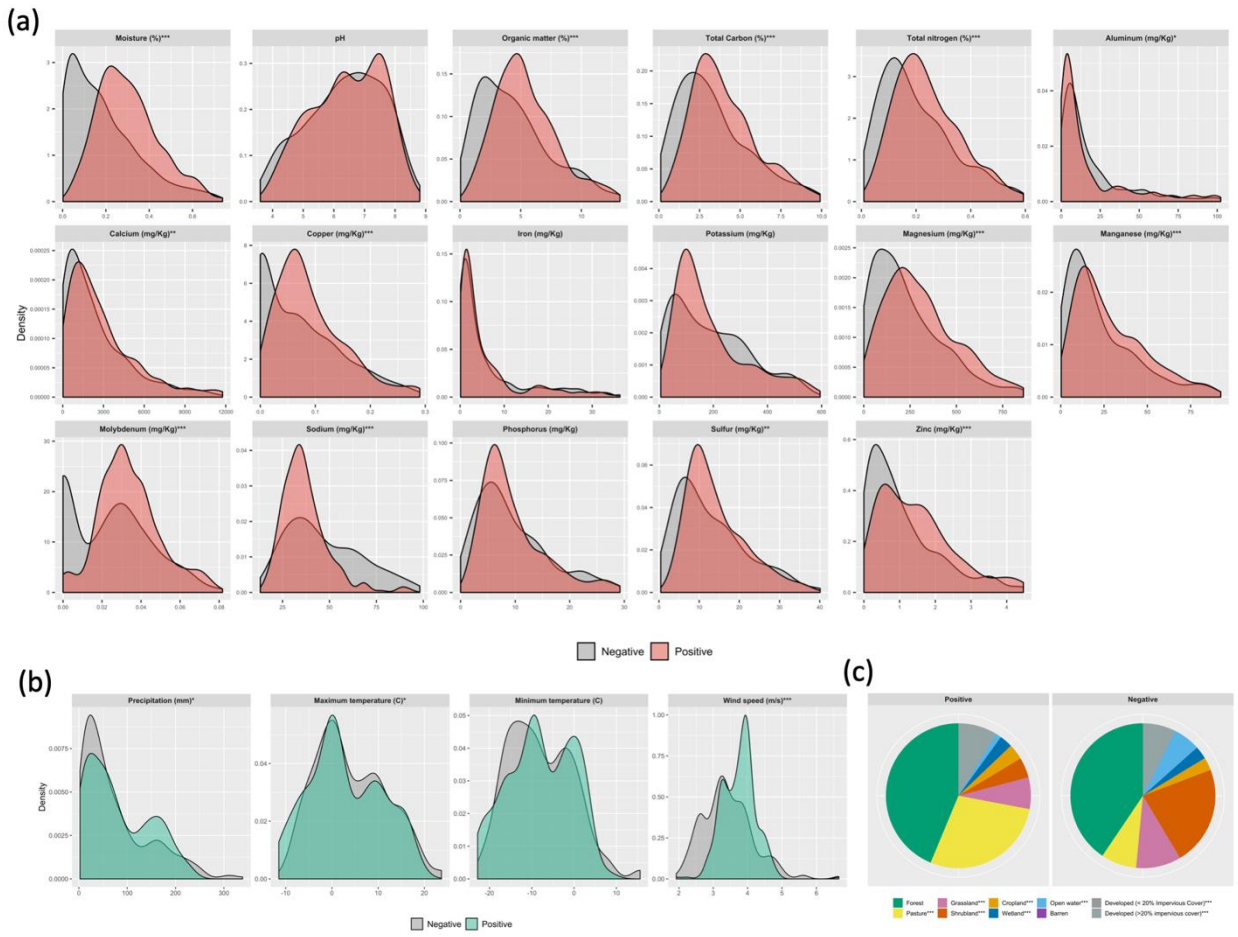

Figure 6. The difference of abiotic variables – (a) soil property, (b) climate, (c) land-use – between positive samples and negative samples. $p < 0.001$ after FDR correction using Benjamini and Hochberg method in Mann-Whitney test are marked as "***"; $p < 0.01$ are marked as "**"; $p < 0.05$ are marked as "*".

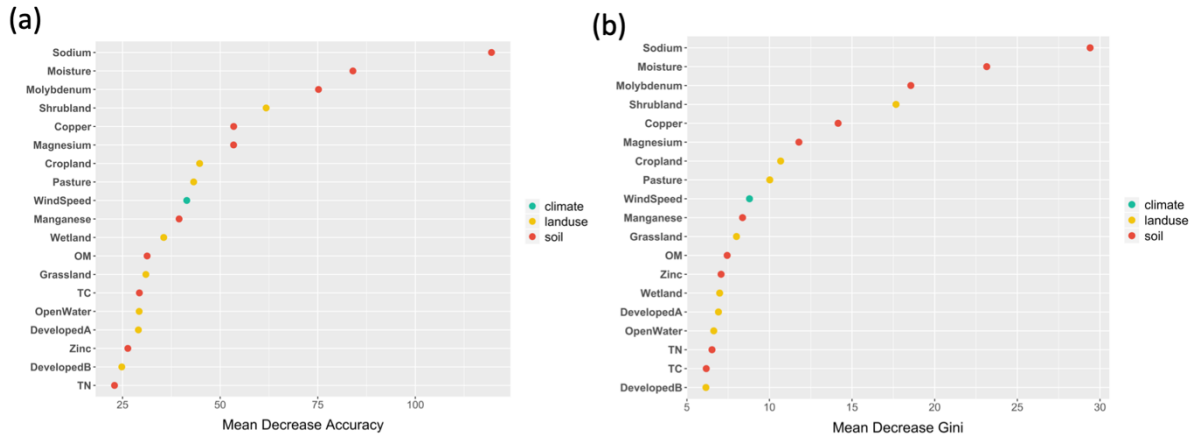

Figure 7. Random Forest model variable importance based on (a) Mean Decrease Accuracy index and (b) Mean Decrease Gini index. Data sample was split into development and validation samples using a probability of 0.7 and 0.3. The number of decision trees was 5000. The accuracy of development sample and validation sample was 0.99 and 0.84, respectively. All 19 variables are sorted in ascending order according to their median importance in the model.

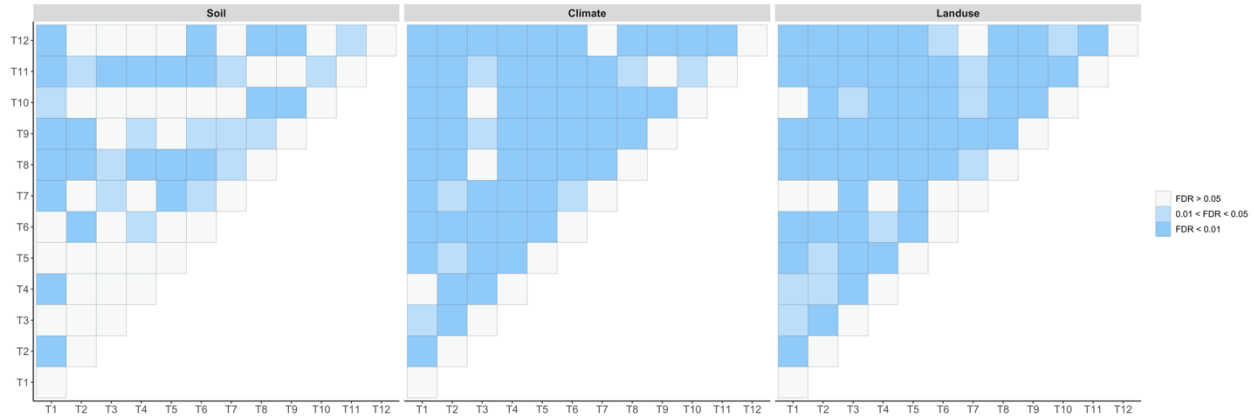

Figure 8. PERMANOVA *post hoc* results for (a) soil property, (b) climate, and (c) land-use. T1: *LM* lineage III; T2: *LM* lineage II; T3: *LM* lineage I; T4: *L. marthii*; T5: unclassified *L. marthii*; T6: *L. innocua*; T7: unclassified *L. innocua*; T8: *L. welshimeri*; T9: *L. seeligeri* lineage I; T10: *L. seeligeri* lineage II; T11: unclassified *L. seeligeri*; T12: *L. booriae*. $p < 0.01$ after FDR correction using Benjamini and Hochberg are indicated by dark blue; $0.01 < p < 0.05$ after FDR correction using Benjamini and Hochberg are indicated by light blue; $p > 0.05$ after FDR correction using Benjamini and Hochberg are indicated by white.

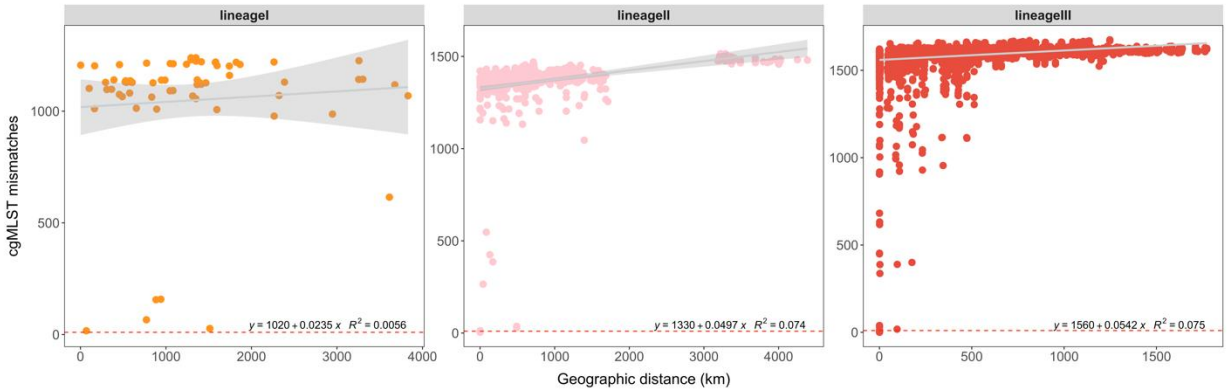

Figure 9. Linear regression between geographic distance and cgMLST mismatches for *LM* (a) lineage I, (b) lineage II, and (c) lineage III. Regression line is reported as well as the formula. R^2 is the determination coefficient.

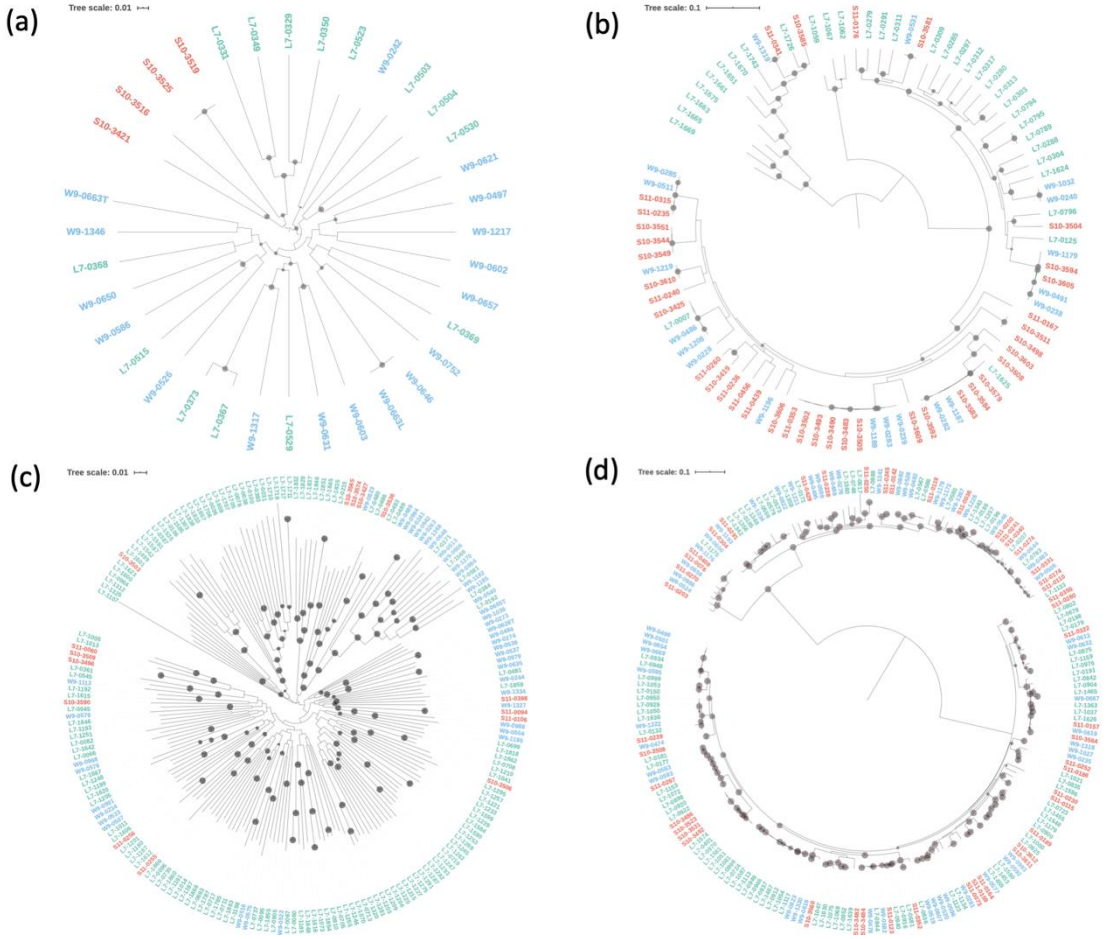

Figure 10. Phylogenetic trees of (a) *L. marthii,* (b) *L. innocua,* (c) *L. welshimeri*, and (d) *L. seeligeri* from non-food and food associated environments based on core SNPs identified by kSNP3. Bootstrap values >0.8 are indicated by grey symbol. Trees were rooted by mid-point. Isolates from soil are marked green; isolates from agricultural water are marked blue; isolates from processing facilities are marked red.

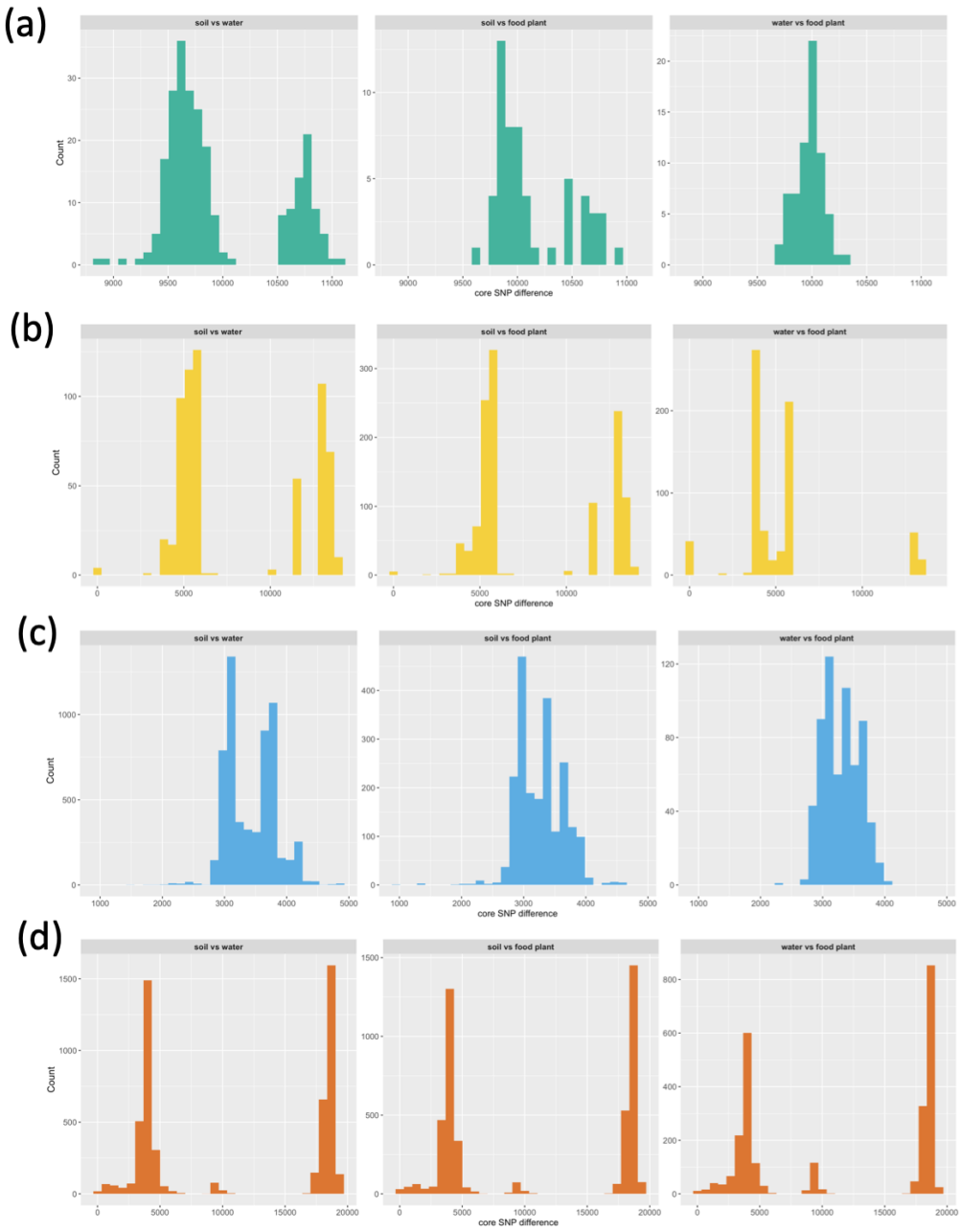

Figure 11. Histogram of core SNP differences for (a) *L. marthii,* (b) *L. innocua,* (c) *L. welshimeri*, and (d) *L. seeligeri* from soil and agricultural water, from soil and processing plant, and from agricultural water and processing plant.

Table 1. Sampling units

Sampling unit	Number	Description
Grids	40	Major areas based on a longitude and latitude grid.
Areas	5 areas/grid; areas within one grid should be >20 km apart	Natural areas (e.g. national/state parks, forests, preserves, wildlife management areas). Cornell suggests areas to each sample collector, but sample collectors can provide input on alternative areas. One sampling kit provided for each area.
Sites	5 sites/area; sites within one area should be >0.2 km apart	Places within an area. Cornell will suggest sites within each area to each sample collector, but sample collectors can provide input on alternative sites based on accessibility.
Points	3 points/site; points within one site should be about 20 feet apart	Better to find 3 points that look similar (e.g., similar type of soil, plant cover). Collect 1 subsample of topsoil (0–8 inches) at each of the 3 sampling points; these 3 subsamples will be pooled for one composite sample per site.

Table 2. cgMLST mismatch between soil *LM* and clinical *LM*

Soil isolate	Clinical isolate	LM lineage	cgMLST mismatch	State of soil LM	Sampling time of soil LM	State of clinical LM	Sampling time of clinical LM
L7-1775	02-2449	I	12	MN	2018-10	missing	missing
L7-1173	PNUSAL000256	I	10	NY	2018-09	missing	2013-08
L7-0846	PNUSAL004600	I	15	OH	2018-08	missing	2018-11
L7-0745	PNUSAL003744	I	16	OR	2018-07	missing	missing
L7-0777	PNUSAL000987	II	12	MA	2019-08	missing	missing
L7-0487	PNUSAL004070	II	17	DE	2018-06	missing	2018-06

Table 3. core SNP and hqSNP differences between *L. seeligeri* from non-food and food associated environments

soil isolate	food associated isolate	core SNP difference	hqSNP difference
L7-0171	W9_1211	2	14
L7-0177	W9_0583	6	23
L7-0181	W9_0583	8	25
L7-0567	W9_0482	5	13
L7-0586	W9_0482	5	16
L7-1175	W9_1176	1	7
L7-0567	S11_0119	2	10
L7-0586	S11_0119	2	13
L7-0925	S10_3611	7	40
L7-0925	S10_3612	7	40
L7-1175	S11_0076	1	11
L7-1175	S11_0408	2	7