Disentangling the impact of abiotic and biotic environmental factors and dispersal dynamics on bacterial pangenome fluidity

Ying-Xian Goh[1,2,3], FNU Hardeep[4], Hailong Zhang[5], Jingqiu Liao[1,2,3*]

[1] Department of Civil and Environmental Engineering, Virginia Tech, Blacksburg, VA 24061, United States
[2] Center for Emerging, Zoonotic, and Arthropod-Borne Pathogens, Virginia Tech, Blacksburg, VA 24061, United States
[3] Global Change Center, Virginia Tech, Blacksburg, VA 24061, United States
[4] Department of Computer Science, Virginia Tech, Blacksburg, VA 24061, United States
[5] Department of Business Information Technology, Virginia Tech, Blacksburg, VA 24061, United States

* Corresponding author: Jingqiu Liao, Email: liaoj@vt.edu

Running title: Ecological drivers of pangenome fluidity

ABSTRACT

Understanding how bacterial pangenomes originate and evolve is crucial for predicting evolutionary trajectories and uncovering ecological interactions. Bacterial pangenome fluidity has been attributed to adaptive evolution, yet its specific ecological drivers remain to be elucidated. *Listeria monocytogenes* (*Lm*), a foodborne pathogen causing fatal listeriosis, serves as an ideal model for studying the ecological mechanisms underlying pangenome fluidity due to its high evolutionary divergence, ecological versatility, and public health relevance. Through pangenome analysis of 177 *Lm* isolates representing three evolutionary lineages (I, II, and III) that we isolated from soil samples systematically collected across the United States, we found that the substantial genome variation among *Lm* genomes was strongly associated with climatic factors (e.g. precipitation and temperature), soil properties (e.g. aluminum, pH, and molybdenum), and bacterial community composition, particularly Nitrospirae, Planctomycetes, Acidobacteria, and Cyanobacteria. These factors exerted strong selective pressure across many gene functions, with pronounced effects on genes involved in cell envelope synthesis, defense mechanisms, and replication, recombination, and repair. Among *Lm* lineages, distinct pangenome properties were observed. Lineage III exhibited a highly fluid pangenome, which was attributed to local adaptation to nutrient-limited conditions and strong dispersal limitation. In contrast, lineage I maintained a conserved pangenome, likely due to frequent homogenizing dispersal. Informed by this dispersal pattern, we identified an increased risk of soil-to-human transmission in lineage I, evidenced by epidemiological links between three soil isolates and 17 clinical isolates. Collectively, these findings highlight the combined roles of environmental selection and dispersal in governing the adaptive evolution of bacterial pangenomes.

INTRODUCTION

Pangenome consists of core genes (i.e., genes present in all individuals) and accessory genes (i.e., genes not shared by all individuals), which arise from gene gain and loss through evolutionary processes such as horizontal gene transfer and pseudogene formation (Nowell et al. 2014; Thomas and Nielsen 2005; Albalat and Cañestro 2016). Bacterial species vary in pangenome fluidity. Some species, like *Escherichia coli*, have an open pangenome, where the number of genes increases substantially with each newly sequenced genome, while others, such as *Bacillus anthracis*, exhibit a closed pangenome, where the gene count remains stable despite the addition of more genomes (McInerney et al. 2017; Rouli et al. 2015; Brockhurst et al. 2019). The evolutionary mechanisms underlying bacterial pangenome fluidity remain a subject of debate in this field. While some argue that pangenome variation results from neutral evolution in which genetic drift randomly changes the frequency of alleles, our previous study and others suggest it is primarily caused by adaptation (Brockhurst et al. 2019; Cummins et al. 2022; Liao et al. 2021). Adaptive evolution of bacterial pangenomes can be triggered by various environmental pressures that bacteria need to cope with, both abiotic, such as acidity, aridity, heat, salinity, aluminum toxicity, and pH (Simonsen 2022; Conrad et al. 2022; Wang et al. 2023; Ma et al. 2024), and biotic, such as bacterial mutualism and competition (Martinson et al. 2023; Ma et al. 2024) and phage predation (Rodriguez-Valera et al. 2009). For example, changes in salinity were found to influence the abundance of accessory genes related to osmoregulation in *Salinibacter ruber* (Conrad et al. 2022). Also, mutualism between *Salmonella enterica* and *Methylorubrum extorquens* in co-culture has been shown to reduce the impact of loss-of-function mutations and selectively maintain nitrogen uptake genes in *S. enterica* compared to its monoculture (Martinson et al. 2023). In addition to environmental selection, pangenome fluidity can also be influenced by dispersal. While dispersal can expose bacteria to new environments and promote the acquisition of novel genes (Vogwill et al. 2010; Garant et al. 2007), pangenome diversity can be constrained by homogenizing dispersal, where the continuous movement of similar genetic material across populations reduces differentiation that would otherwise emerge from localized environmental pressures (Richardson et al. 2014; Bourne et al. 2014; Crits-Christoph et al. 2020). Linking pangenome properties to these ecological factors can shed light on the mechanisms underlying bacterial evolution and enhance its predictability under environmental changes. However, the ecological drivers of bacterial pangenome fluidity remain to be elucidated due to a lack of in-depth investigations that integrate genomics with paired abiotic and biotic environmental data.

Listeria monocytogenes (*Lm*) is characterized by high genomic diversification, ecological versatility, and public health relevance, which make it an ideal model for studying the ecological mechanisms governing bacterial pangenome evolution. *Lm* is a Gram-positive, motile, non-spore-forming facultative bacterial species commonly found as a saprophyte in soils (Freitag et al. 2009; Orsi and Wiedmann 2016). *Lm* is also a foodborne pathogen that causes rare but fatal

listeriosis affecting vulnerable populations (i.e., pregnant women, the elderly, infants, and immunocompromised individuals), with a fatality rate of 20-30%, hospitalization rate of 94%, and an annual illness cost of USD 2.6 billion (Hoffmann et al. 2012; US Food and Drug Administration 2020). *Lm* can thrive and transmit across diverse habitats (e.g., soil, water, animals, produce, food processing facilities, and hospital settings) (Liao et al. 2021, 2023; US Food and Drug Administration 2020; Koopmans et al. 2023; Freitag et al. 2009). It has formed four distinct evolutionary lineages (designated as I, II, III, and IV) (Orsi et al. 2011) characterized by substantial intraspecific gene content variation (den Bakker et al. 2013, 2012; Wang et al. 2019) and unique ecological roles (Linke et al. 2014; Liao et al. 2023). *Lm* lineage I, II, and III are commonly detected in human clinical cases, food-related sources, and soil environments, respectively, while lineage IV is rare and has been detected in ruminants (Orsi et al. 2011; Liao et al. 2021, 2023). It is known that *Lm* genomes undergo adaptive evolution (Liao et al. 2021; Ryan et al. 2010; Harter et al. 2017; Maury et al. 2019; Moura et al. 2016). For example, *Lm* has developed stress survival islets (SSI) 1 and 2 to cope with environmental stress, with SSI-1 responsible for tolerating high salt and bile salt concentrations and low pH (Ryan et al. 2010), and SSI-2 for tolerance of alkaline and oxidative stress conditions (Harter et al. 2017). Also, in food processing environments where benzalkonium chloride is commonly used as a surface disinfectant, *Lm* strains have been found to acquire tolerance genes to withstand its effects (Maury et al. 2019; Moura et al. 2016). These individual studies, however, only provide a snapshot of how environmental factors influence particular genes in *Lm*, with a focus on food-associated environments. The systematic effects of abiotic and biotic factors shaping the pangenome fluidity of *Lm* inhabiting natural environments remain largely unknown.

Here, we conducted a comprehensive pangenome analysis of 177 *Lm* isolates obtained from 1,004 soil samples that we previously systematically collected from natural environments across the United States (US) (Liao et al. 2021). These isolates represent three lineages, I, II, and III. We examined the relationships between the pangenome variation of *Lm* and 34 abiotic variables capturing geolocation, soil properties, climate, and surrounding land use, as well as bacterial community composition characterized by 16S rRNA gene amplicon sequencing. Using a suite of statistical analyses, including machine learning, we identified strong relationships between climate, soil properties, and bacterial community composition, particularly Nitrospirae, Planctomycetes, Acidobacteria, and Cyanobacteria, and pangenome variation at multiple functional categories. Functional enrichment analysis revealed that abiotic and biotic pressures jointly target genes involved in cell envelope synthesis and regulation, with abiotic and biotic factors additionally acting on genes related to defense mechanisms and replication, recombination, and repair, respectively. We also identified distinct pangenome properties across *Lm* lineages. The pangenome of lineage III is highly fluid, but its genome size appears to be streamlined, which we attribute to local adaptation to nutrient-limited conditions and strong dispersal limitation. In contrast, the pangenome of lineage I is conserved, which is likely contributed by frequent homogenizing dispersal across the US. Informed by this observation, we

identified a high risk of transmission from soil to humans in lineage I through comparative genomic analysis of soil and clinical isolates. Overall, this study provides insights into how selective pressures triggered by abiotic factors and bacterial communities and dispersal dynamics can contribute to bacterial pangenome fluidity.

RESULTS

Associations between abiotic environmental factors and *Lm* pangenome

For the 177 *Lm* isolates included in this study, a total of 5,873 orthologous genes were identified, with 2,324 and 3,549 classified as core and accessory genes, respectively (**Supplemental Fig. S1**). The gene richness per genome ranged from 2627 to 3082 (**Supplemental Fig. S1**). To identify key abiotic drivers of the pangenome composition of *Lm*, we performed a series of statistical analyses assessing the relationships between 34 abiotic environmental variables for geolocation, soil properties, climate, and surrounding land use, and gene richness as well as gene presence/absence patterns. Variation partitioning analysis (VPA), combined with a permutation test (see Methods), revealed that abiotic factors altogether significantly explained 14.24% of the gene richness variation (one-sided permutation $P = 0.01$; **Fig. 1A**). Among them, soil property and climatic variables were found to be most influential, accumulatively explaining 10.17% and 9.05% of the variation with other variables, respectively, followed by geolocation (8.79%) and surrounding land use (6.36%; **Fig. 1A**). Of note, the interplay of soil properties and climate explained 5.16% of the variation (**Fig. 1A**). To further assess the relationships between abiotic factors and gene richness at the functional level, VPA stratified by Clusters of Orthologous Groups (COGs) was conducted. Results showed that among the 18 COGs, abiotic factors significantly explained the variation of gene richness for 17 COGs (adjusted one-sided permutation $P < 0.05$ for all), with G (carbohydrate transport and metabolism) and K (transcription) exhibiting over 50% of the variation explained (**Fig. 1B**).

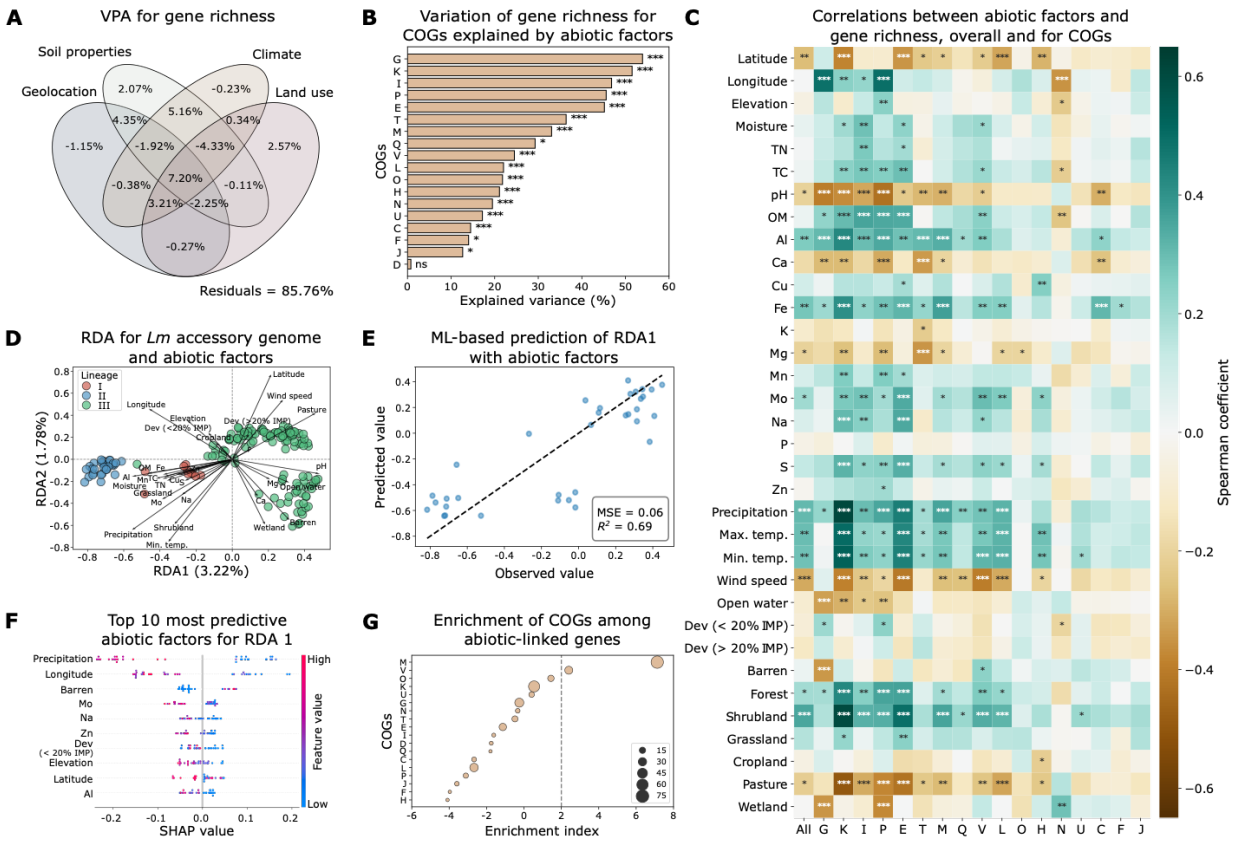


Figure 1. Pangenome fluidity of *Lm* is associated with abiotic factors.

(A) Venn diagram of variation partitioning analysis (VPA) showing the variation of gene richness explained by geolocation, soil properties, climate, and surrounding land use. Residuals represent unexplained variation. A negative explained variation may be caused by collinearity between predictors and overparameterization of the model. **(B)** Variation of gene richness for COGs explained by abiotic factors in VPA. Statistical significance was assessed using a one-sided permutation test. **(C)** Spearman's correlation between abiotic factors and gene richness, overall and for significant COGs identified in VPA. Positive and negative correlations are indicated by green and brown, respectively. **(D)** Redundancy analysis (RDA) biplot illustrating the relationships between the *Lm* accessory genome and 31 abiotic variables selected by Lasso regularization. Points represent individual genomes, color-coded by lineages. The length and direction of the arrows indicate the relative importance and correlation of each abiotic variable with the ordination axes. The first (RDA1) and second (RDA2) ordination axes explain 3.22% and 1.78% of the variation of the *Lm* accessory genome explained by selected abiotic factors, respectively. **(E)** Prediction of RDA1 axis values with abiotic factors using the LightGBM machine learning (ML) model. MSE, mean squared error; R^2, coefficient of determination. The dashed line represents the line of perfect agreement (y = x) where predicted values would exactly match observed values. **(F)** The top ten most predictive abiotic factors for RDA1 (SHAP-based; X axis), sorted by descending importance. SHAP values indicate the impact of features on ML model output. **(G)** Enrichment of COGs among abiotic-linked genes. An enrichment index greater than two (grey dashed line) indicates significant overrepresentation ($P < 0.05$). Circle size is proportional to the number of genes annotated to each COG. For **(B)** and **(C)**, significance levels are denoted by "*", "**", "***", and "ns" for adjusted $P < 0.05$, < 0.01, < 0.001, and > 0.05 (not significant), respectively. Abbreviations of COGs for **(B)** and **(G),** and abiotic factors for **(C)**, **(D)**, and **(F)** are described in Methods.

To assess the relationships between individual abiotic factors and gene richness, Spearman's correlation analysis was performed for overall gene richness as well as that for each COG. For overall gene richness, precipitation, temperature, soil aluminum, iron, and surrounding shrubland coverage were significantly positively correlated with gene richness, while wind speed, magnesium, pH, and surrounding pasture coverage showed a significant negative correlation (adjusted Spearman's $P < 0.05$ for all; **Fig. 1C**). At the functional level, these abiotic variables were also significantly correlated with gene richness for over half of the 17 significant COGs identified in VPA (adjusted $P < 0.05$ for all; **Fig. 1C; Supplemental Fig. S2**), suggesting a universal effect on the gene content of *Lm* across multiple functions. In contrast, some abiotic variables tend to have a function-specific effect on gene richness. For example, potassium and zinc were only significantly correlated with gene richness for COGs T (signal transduction mechanisms) and P (inorganic ion transport and metabolism), respectively (adjusted $P < 0.05$ for both; **Fig. 1C**).

In addition to gene richness, we assessed the relationships between abiotic variables and the presence/absence patterns of accessory genes. Redundancy analysis (RDA) of 31 out of 34 abiotic variables selected by Lasso regularization showed that 3.22% and 1.78% of the variation in the first (RDA1) and second (RDA2) ordination axis of the accessory gene content, respectively, was explained by these selected variables, with precipitation, minimum temperature, wind speed, aluminum, pH, molybdenum, latitude, longitude, and surrounding shrubland, pasture, barren, and wetland coverage exerting a strong effect (**Fig. 1D**). As the relationships between abiotic variables and accessory gene variance may not be captured by the linear constraints of RDA, we further employed machine learning to model their relationships (see Methods). Results show that with LightGBM and random forests, 69% and 84% of the variation in RDA1 and RDA2 was explained by these selected abiotic variables, respectively (**Fig. 1E; Supplemental Fig. S3A**). Based on Shapley Additive exPlanations (SHAP) analysis (Lundberg and Lee 2017), precipitation, minimum temperature, aluminum, molybdenum, latitude, longitude, and surrounding coverage of shrubland and barren, which were found to have a strong effect in the RDA biplot (**Fig. 1D**), were also among the top 10 most predictive abiotic factors for RDA1 and/or RDA2 (**Fig. 1F; Supplemental Fig. S3B**). Analyses on gene richness and accessory genomes altogether suggest that the pangenome fluidity of *Lm* is strongly shaped by climatic variables (e.g., precipitation and temperature), soil properties (e.g., aluminum, pH, and molybdenum), and geolocation.

To assess the associations between abiotic factors and individual accessory genes, we compared the difference of abiotic factors for *Lm* genomes with and without a given accessory gene. A total of 803 accessory genes were found to be significantly associated with at least one abiotic factor, which we defined as "abiotic-linked genes" (adjusted Mann-Whitney U $P < 0.05$ for all; **Supplemental Table S1**). Functional enrichment analysis (see Methods) showed that COGs M (cell wall/membrane/envelope biogenesis) and V (defense mechanisms) were

significantly enriched among these abiotic-linked genes (**Fig. 1G**). These results along with gene richness analysis stratified by COGs suggest that abiotic factors play a critical role in structuring the gene content of *Lm* across multiple functions, with pronounced effects acting on genes involved in cell envelope formation and defense mechanisms.

Associations between bacterial community composition and *Lm* pangenome

Lm does not exist in isolation within the soil environment; instead, it coexists within communities alongside other bacteria. Among the soil samples positive for *Lm*, a total of 28 phyla were identified. To test the contribution of bacterial community composition (also referred to as "biotic factors") to the variation of *Lm* pangenome, VPA was performed on the gene richness and the relative abundance of bacterial phyla along with abiotic factors. Bacterial phylum composition alone and cumulatively with abiotic factors significantly contributed to 13.25% and 19.87% of the variation of gene richness, respectively (one-sided permutation $P < 0.001$; **Fig. 2A**). Stratifying by COGs, the variation of gene richness for 14 COGs was significantly explained by both abiotic factors and bacterial phylum composition (adjusted one-sided permutation $P < 0.05$ for all; **Fig. 2B**). Among these COGs, G (carbohydrate transport and metabolism), I (lipid transport and metabolism), P (inorganic ion transport and metabolism), and K (transcription) had over 50% of their variation explained, indicating a significant impact of abiotic and biotic factors on these gene functions (**Fig. 2B**). Of note, COGs L (replication, recombination, and repair), J (translation, ribosomal structure, and biogenesis), and V (defense mechanisms) showed greater variation of gene richness explained by biotic factors than abiotic factors, indicating higher importance of bacterial community composition impacting these functions (**Fig. 2A**).

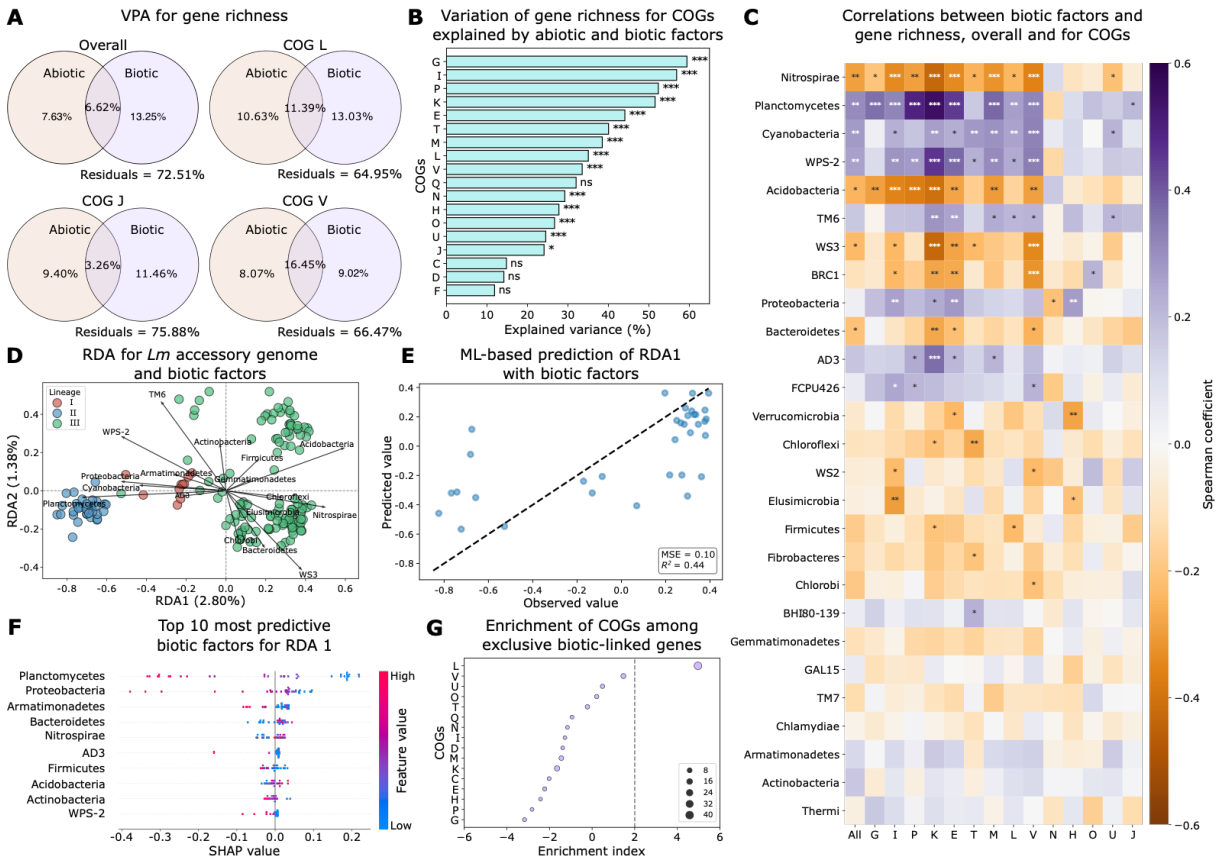


Figure 2. Pangenome fluidity of *Lm* is associated with bacterial community composition. **(A)** Venn diagram of VPA showing the variation of gene richness explained by abiotic and biotic factors. Biotic factors are represented by bacterial community composition at the phylum level. COGs where biotic factors explain more variation than abiotic factors are highlighted. Residuals represent unexplained variation. **(B)** Variation of gene richness for COGs explained by abiotic and biotic factors in VPA. Statistical significance was assessed using a one-sided permutation test. **(C)** Spearman's correlation between the relative abundance of bacterial phyla and gene richness, overall and for significant COGs identified in VPA. Positive and negative correlations are indicated by purple and orange, respectively. **(D)** RDA biplot illustrating the relationships between the *Lm* accessory genome and 17 phyla selected by Lasso regularization. Points represent individual genomes, color-coded by lineages. The length and direction of the arrows indicate the relative importance and correlation of each phylum with the ordination axes. RDA1 and RDA2 account for 2.80% and 1.38% of the variation of the *Lm* accessory genome explained by selected phyla, respectively. **(E)** Prediction of RDA1 axis values with relative abundance of bacterial phyla using random forests. MSE, mean squared error; R^2, coefficient of determination. The dashed line represents the line of perfect agreement (y = x) where predicted values would exactly match observed values. **(F)** The top ten most predictive biotic factors for RDA1 (SHAP-based; X axis), sorted by descending importance. SHAP values indicate the impact of features on ML model output. **(G)** Enrichment of COGs among exclusive biotic-linked genes. An enrichment index greater than two (indicated by the grey dashed line) signifies significant enrichment ($P < 0.05$). Circle size is proportional to the number of genes annotated to each COG. Abbreviations of COGs are described in Methods. For

258 **(B)** and **(C)**, significance levels are denoted by "*", "**", "***", and "ns" for adjusted $P < 0.05$, < 0.01,
259 < 0.001, and > 0.05 (not significant), respectively.
260
261 To pinpoint specific bacterial phyla associated with the gene richness of *Lm*, Spearman's
262 correlation analysis was performed on the relative abundance of phyla and overall gene richness
263 as well as that for each COG. Results showed that Planctomycetes, Cyanobacteria and WPS-2
264 were significantly positively correlated with gene richness, while Nitrospirae, Acidobacteria,
265 WS3, and Bacteroidetes showed a significant negative correlation (adjusted Spearman's $P <$
266 0.05 for all; **Fig. 2C**). At the functional level, Nitrospirae, Planctomycetes, Cyanobacteria, WPS-
267 2, and Acidobacteria were significantly correlated with gene richness for over half of the 14
268 significant COGs identified in VPA (adjusted $P < 0.05$ for all; **Fig. 2C; Supplemental Fig. S4**),
269 suggesting a universal effect on the gene content of *Lm* across multiple functions. In contrast,
270 some phyla tend to have a function-specific effect on gene richness. For example, Fibrobacteres
271 and Chlorobi were only significantly correlated with gene richness for COGs T (signal
272 transduction mechanisms) and V (defense mechanisms), respectively (adjusted $P < 0.05$ for
273 both; **Fig. 2C**).
274
275 In addition to gene richness, we further examined the relationships between bacterial community
276 composition and the presence/absence patterns of accessory genes. RDA of 17 out of 27 phyla
277 selected by Lasso regularization showed that 2.80% and 1.38% of the variation in RDA1 and
278 RDA2 of the accessory gene content, respectively, was explained by these selected phyla, with
279 TM6, Planctomycetes, WS3, WPS-2, Acidobacteria, Proteobacteria, Nitrospirae, Bacteroidetes,
280 Cyanobacteria, and Chlorobi exerting a strong effect (**Fig. 2D**). Machine learning models further
281 enhanced the prediction, with 44% and 42% of the variation in RDA1 and RDA2 being explained
282 by phyla using random forests and gradient boosting, respectively (**Fig. 2E; Supplemental Fig.**
283 **S5A**). Based on SHAP analysis (Lundberg and Lee 2017), TM6, Planctomycetes, WS3, WPS-
284 2, Acidobacteria, Proteobacteria, Nitrospirae, Bacteroidetes, and Cyanobacteria which were
285 found to have a strong effect in the RDA biplot (**Fig. 2D**), were also among the top 10 most
286 predictive phyla for RDA1 and/or RDA2 (**Fig. 2F; Supplemental Fig. S5B**). Analyses on gene
287 richness and accessory genomes altogether suggest that in addition to abiotic factors, bacterial
288 community composition, particularly Nitrospirae, Planctomycetes, Acidobacteria, and
289 Cyanobacteria, play an important role in pangenome fluidity of *Lm*.
290
291 To assess the associations between biotic factors and individual accessory genes, we compared
292 the relative abundance of phyla for *Lm* genomes with and without a given accessory gene. A
293 total of 807 accessory genes were found to be significantly associated with at least one phylum,
294 referred to as "biotic-linked genes" (adjusted Mann-Whitney U $P < 0.05$ for all; **Supplemental**
295 **Table S2**). Functional enrichment analysis showed that these biotic-linked genes were
296 significantly enriched for cell wall/membrane/envelope biogenesis (M) (**Supplemental Fig.**
297 **S6A**). Comparing these 807 biotic-linked genes to the 803 abiotic-linked genes, 299 genes
298 (27.0%) and 303 genes (27.4%) were exclusively linked to abiotic factors and biotic factors,

respectively (**Supplemental Fig. S6B**). Functional enrichment analysis showed that COGs M (cell wall/membrane/envelope biogenesis) and V (defense mechanisms) were significantly enriched among genes exclusively linked to abiotic factors (**Supplemental Fig. S6C**), while L (replication, recombination, and repair) was significantly enriched among genes exclusively linked to biotic factors (**Fig. 2G**), consistent with the VPA result that biotic factors were found to be more important for gene richness involved in this function than abiotic factors (**Fig. 2A**). These results along with gene richness analysis stratified by COGs suggest that bacterial community composition influence the gene content of *Lm* across multiple functions, with pronounced effects on genes involved in cell envelope formation jointly with abiotic factors and uniquely in replication, recombination, and repair.

Pangenome variation among *Lm* lineages

The 177 *Lm* isolates were classified into three lineages, lineage I (12 isolates), II (29 isolates), and III (126 isolates). Geographically, lineages I (found in the northern and central regions) and II (found along the eastern seaboard and parts of the west) were more widely distributed, whereas lineage III formed a geographic cluster in the central and eastern areas (**Fig. 3A**). Among the 34 abiotic variables, 22 were significantly different among the three lineages (adjusted Kruskal-Wallis $P < 0.05$ for all; **Supplemental Fig. S7**). Of note, lineage III appears to be more capable of surviving in nutrient-limited conditions compared to lineages I and II, as its habitats had significantly lower levels of organic matter, total nitrogen, sodium, and sulfur (adjusted Mann-Whitney U $P < 0.05$ for all; **Supplemental Fig. S7**). In the multidimensional scaling (MDS) analysis based on the 34 abiotic variables, genomes formed significantly distinct clusters by lineages (permutational multivariate analysis of variance, PERMANOVA, $P = 0.001$, pseudo-$F = 6.703$; **Fig. 3B**), which suggests that *Lm* lineages inhabit significantly different abiotic conditions.

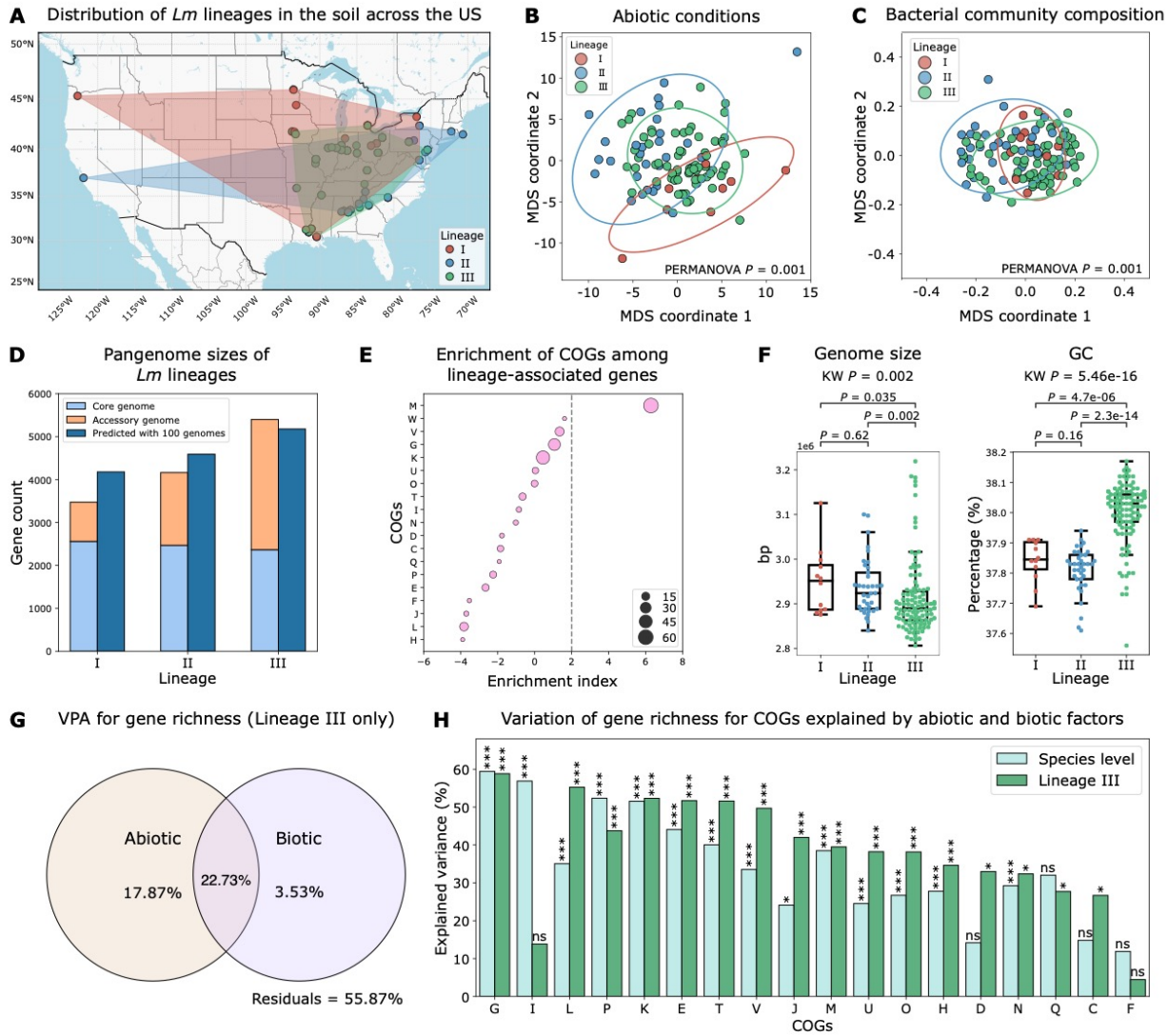


Figure 3. Pangenome variation is substantial among *Lm* lineages.
(A) Map showing the distribution of *Lm* lineages in soils across the US. Points represent *Lm*-positive samples, color-coded by lineages. Polygons indicate the habitat range of each lineage. This map was adapted from **Extended Data Fig. 5** in Liao et al., 2021.(Liao et al. 2021) **(B-C)** Multidimensional scaling (MDS) analysis for *Lm* lineages based on **(B)** abiotic conditions (pseudo-F = 6.703) and **(C)** bacterial community composition using weighted UniFrac distances based on OTUs (pseudo-F = 8.829). Points represent *Lm* genomes color-coded by lineages. PERMANOVA P = 0.001 indicates significant differences in abiotic and biotic conditions among lineages. Ellipse size represents two times the standard deviation from the mean. **(D)** Pangenome sizes of *Lm* lineages, stratified into core and accessory genome sizes, and predicted pangenome sizes based on 100 genomes per lineage. Pangenome size predictions were made using the power law function $cN\gamma$ estimated in **Extended Data Fig. 7** in Liao et al., 2021.(Liao et al. 2021) **(E)** Enrichment of COGs among lineage-associated genes. An enrichment index greater than two (indicated by the grey dashed line) signifies significant enrichment (P < 0.05). The size of the circles is proportional to the number of genes annotated to each COG. COG abbreviations are described in

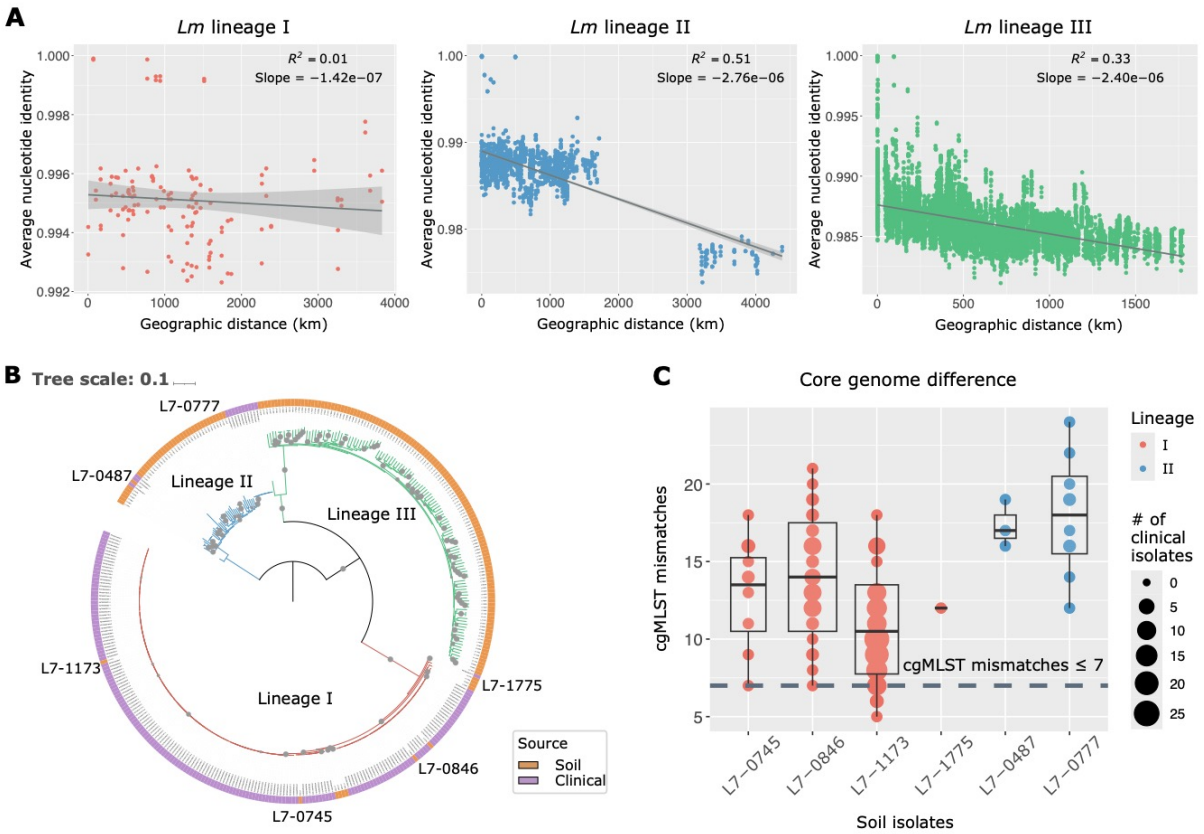
Methods. **(F)** Genome size and GC content compared among *Lm* lineages. Box plots display the interquartile range (IQR) with the median indicated as a line and whiskers extending to 1.5 times the IQR. Adjusted Kruskal-Wallis (KW) *P* value and adjusted two-sided Mann-Whitney *U P* values are annotated for each boxplot. **(G)** Venn diagram of VPA showing the variation of gene richness for lineage III explained by abiotic and biotic factors. Residuals represent unexplained variation. **(H)** Variation of gene richness for COGs explained by abiotic and biotic factors at the species level (cyan) and within lineage III (green). Significance levels are denoted by "*", "**", "***", and "ns" for adjusted $P < 0.05$, < 0.01, < 0.001, and > 0.05 (not significant), respectively, and "ns" in one-sided permutation tests for VPA.

Similarly, we also identified significant differences in the bacterial community composition compared among *Lm* lineages. Of the 28 phyla, Nitrospirae, Planctomycetes, Acidobacteria, and Proteobacteria showed significantly different relative abundance among the three lineages (adjusted Kruskal-Wallis $P < 0.05$ for all; **Supplemental Fig. S8**). Compared to other lineages, the relative abundance of Planctomycetes and Proteobacteria was significantly higher in lineage II and lower in lineage III, respectively (adjusted Mann-Whitney *U P* < 0.05 for all). Acidobacteria and Nitrospirae were significantly more abundant in lineage III than in lineage II (adjusted P = 2.0e-06 and 1.3e-06, respectively; **Supplemental Fig. S8**). MDS analysis showed that genomes formed significantly distinct clusters by lineages based on the beta diversity of bacterial communities represented by weighted UniFrac distances (pseudo-F = 5.829; PERMANOVA P = 0.001; **Fig. 3C**). Overall, these results suggest that *Lm* lineages occupy significant different abiotic and biotic environmental conditions, which may facilitate adaptive genomic variation in *Lm*.

Indeed, we observed large differences in the gene content among the three *Lm* lineages. While their core genome sizes were similar (2,560, 2,468, and 2,366 core genes for lineage I, II, and III, respectively), their accessory genome sizes were substantially different (913, 1,695, and 3,034 accessory genes for lineage I, II, and III, respectively; **Fig. 3D**). Using the power law function estimated from the pangenome curves for these isolates (**Extended Data Fig. 7** in Liao et al., 2021(Liao et al. 2021)), we predicted the pangenome size for each lineage given 100 genomes, which was 4,179, 4,592, and 5,178 genes for lineage I, II, and III, respectively (**Fig. 3D**). Using Fisher's exact tests, we identified 670 accessory genes significantly associated with lineages (adjusted $P < 0.05$ for all; **Supplemental Table S3**), consistent with the RDA result which identified three distinct clusters by lineages based on accessory genes (**Fig. 1D; Fig. 2D**). These lineage-associated genes were significantly enriched for cell wall/membrane/envelope biogenesis (M) based on the functional enrichment analysis (**Fig. 3E**). In addition, we compared the genome size, GC content, virulence factors (*Listeria* pathogenicity island [LIPI]-1, -3, and -4, and internalin [*inl*] genes), SSI 1-2, antibiotic resistance genes (ARGs; *fosX, lin, mprF, norB,* and *sul*), and mobile genetic elements (MGEs), including insertion sequences (IS), transposons, prophages, and plasmids, among lineages. Compared to lineages I and II, lineage III had a significantly smaller genome size (adjusted Mann-Whitney *U P* = 0.035 and 0.0022, respectively, **Fig. 3F**) and higher GC content (adjusted P = 4.7e-06 and 2.3e-14, respectively, **Fig. 3F**). For

381 virulence genes, *prfA* (required for the transcription of LIPI-1 (Behari and Youngman 1998)) and
382 *hly* (encoding the pore-forming cytolysin Listeriolysin O (Behari and Youngman 1998)) in LIPI-1,
383 *inl* genes (encoding cell wall-bound internalins (Bierne et al. 2007)), LIPI-3 genes (encoding the
384 hemolytic and cytotoxic Listeriolysin S [LLS] (Cotter et al. 2008)), and LIPI-4 genes (encoding a
385 cellobiose-family phosphotransferase system (Maury et al. 2016)) all exhibit significant
386 differences in frequency compared among lineages (adjusted Fisher's exact $P < 0.05$ for all;
387 **Supplemental Fig. S9A**). Of note, SSI-2, which confers resistance to higher pH (Harter et al.
388 2017), was significantly overrepresented in lineage III (adjusted Fisher's exact $P = 3.0e-11$;
389 **Supplemental Fig. S9A**), consistent with the observation the habitats of lineage III had
390 significantly higher pH compared to lineage II (adjusted Mann-Whitney *U P* = 5.8e-05;
391 **Supplemental Fig. S7**). In addition, proportions of IS elements and transposons significantly
392 differed among lineages, with lineage II having a significantly higher proportion (Kruskal-Wallis
393 $P = 1.2e-13$ and 9.8e-08, respectively; **Supplemental Fig. S9B**). No significant differences were
394 observed for ARGs (adjusted Fisher's exact test $P > 0.05$ for all; **Supplemental Fig. S9A**),
395 prophages (Kruskal-Wallis $P = 0.88$; **Supplemental Fig. S9B**), and plasmids (Kruskal-Wallis P
396 $= 0.07$; data not shown due to low prevalence).
397
398 To compare the associations between pangenomes and abiotic and biotic factors among *Lm*
399 lineages, we intended to perform VPA on gene richness stratified by lineages. However, none
400 of the variation was explained for lineages I and II likely due to the limited sample sizes ($n = 12$
401 and 29 for lineage I and II, respectively). Therefore, we compared the VPA results for lineage III
402 and the ones at the species level. We found that abiotic and biotic factors jointly explained
403 44.13% of the variation of gene richness for lineage III (one-sided permutation $P = 0.01$; **Fig.**
404 **3G**), which is substantially higher than the 27.49% observed for overall gene richness for *Lm*
405 (**Fig. 2A**). Further, VPA stratified by COGs identified 16 COGs for which variation of gene
406 richness was significantly explained by abiotic and biotic factors. Notably, carbohydrate transport
407 and metabolism (G), replication, recombination, and repair (L), transcription (K), amino acid
408 transport and metabolism (E), and signal transduction mechanisms (T) each exhibited over 50%
409 of the variation explained by these factors (adjusted one-sided permutation $P < 0.05$ for all; **Fig.**
410 **3H**). Among the 18 COGs, 13 COGs (72.2%) showed a higher proportion of variation explained
411 in lineage III compared to the species-level result (**Fig. 3H**). These results suggest that the
412 influence of abiotic and biotic factors on the gene content of *Lm* is driven by lineage III and this
413 lineage appears to undergo strong adaption to local nutrient-limited conditions.
414
415 **Dispersal dynamics of *Lm* lineages**
416
417 Homogenizing dispersal across geographic locations can facilitate evolutionary processes that
418 increase genetic similarity among bacterial populations (e.g., homologous recombination)
419 (Richardson et al. 2014; Bourne et al. 2014; Crits-Christoph et al. 2020). In contrast, dispersal
420 limitation can enhance differentiation in genetic material across populations that emerge from

421 localized environmental adaptation (Bourne et al. 2014; Richardson et al. 2014). To understand
422 the dispersal patterns of *Lm* lineages, we examined the correlations between genetic similarities,
423 measured by average nucleotide identity (ANI), and geographic distances for each lineage. A
424 distance-decay relationship, in which genetic similarities decline with geographic distances,
425 suggests dispersal limitation (Legendre et al. 2005; Bell 2010). Results show that lineage I
426 exhibited a weak distance-decay relation (R^2 = 0.01, slope = 1.42e-07; **Fig. 4A**), suggesting
427 frequent homogenizing dispersal across locations. In contrast, lineages II and III showed a
428 strong distance-decay relationship (R^2 = 0.51 and 0.33, slope = -2.76e-06 and -2.40e-06,
429 respectively; **Fig. 4A**), indicating strong dispersal limitation. These different dispersal patterns
430 could partially explain the more fluid pangenomes of lineages II and III compared to lineage I.
431 The important role of dispersal in pangenome fluidity is also supported by the strong correlations
432 observed between geolocation and variation in gene content (**Fig. 1A; Fig. 1D; Fig. 1F**).



433
434 **Figure 4. Dispersal patterns vary among *Lm* lineages.**
435 **(A)** Linear regression for genetic similarities measured by average nucleotide identity (ANI) and
436 geographic distances for *Lm* lineage I, II, and III. A steeper negative slope of the fitted regression line
437 with a higher R^2 indicates a stronger distance-decay relationship. Shaded areas depict the 95%
438 confidence interval (mean ± 1.96 standard error of mean, SEM) of the linear regression. **(B)** Maximum
439 likelihood phylogenetic tree based on core single nucleotide polymorphisms (SNPs) of 177 soil *Lm*
440 isolates (orange) and 186 closely related clinical isolates (purple), with 1,000 bootstrap replicates.
441 Bootstrap values > 80% are marked by gray circles. The tree is rooted by midpoint, and branches are

442 color-coded by lineages. The six soil isolates with epidemiological links are annotated. **(C)** Core genome
443 multi-locus sequence typing (cgMLST) allelic mismatches between six soil isolates of lineages I (red) and
444 II (blue) and 186 closely related clinical isolates. The horizontal dashed grey line indicates the cutoff (≤ 7
445 cgMLST mismatches) for documented epidemiological links.(Moura et al. 2016) The size of the circles is
446 proportional to the number of clinical isolates closely related to a given soil isolate.
447
448 Given the frequent homogenizing dispersal potential observed in *Lm* lineage I, we hypothesized
449 that this lineage may have a higher risk of transmission from soils to humans. To test this
450 hypothesis, we compared the genomes of our soil isolates with clinical isolates from human
451 infection downloaded from the National Center for Biotechnology Information (NCBI) Pathogen
452 Detection Isolates Browser (National Center for Biotechnology Information 2016). A total of 186
453 clinical isolates were found to be closely related to six soil isolates (< 50 single-nucleotide
454 polymorphisms [SNPs]) (see Method; **Supplemental Figs. S10 and S11; Supplemental Table**
455 **S4**). Among these clinical isolates, 173 and 13 were closely related to four soil isolates of lineage
456 I (L7-0745, L7-0846, L7-1173, and L7-1775) and two soil isolates of lineage II (L7-0487 and L7-
457 0777), respectively (**Fig. 4B; Supplemental Figs. S10 and S11; Supplemental Table S4**).
458 Subsequent analysis using core genome multi-locus sequence typing (cgMLST) identified 17
459 clinical isolates with seven or fewer allelic mismatches with three soil isolates of lineage I (L7-
460 0745, L7-0846, and L7-1173), suggesting epidemiological links (**Fig. 4C**) (Fig. 4C) (Moura et al.
461 2016). Specifically, L7-0745 and L7-0846 each was linked with one clinical isolate, while L7-
462 1173 was linked to 15 clinical isolates (**Supplemental Table S4**). Genomic comparisons,
463 including genome size, GC content, the count of LIPI-1, *inl*, LIPI-3, and LIPI-4 genes, ARGs, and
464 proportions of MGEs (including IS, transposons, and prophages), between these three soil
465 isolates and their corresponding linked clinical isolates further supported the epidemiological
466 links, as no significant differences were identified (Wilcoxon rank-sum $P > 0.05$ for all;
467 **Supplemental Fig. S12**). None of the clinical and soil isolates compared contain any SSI or
468 plasmids. However, it is worth noting that clinical isolates consistently carried more ARG copies
469 than the linked soil isolates (**Supplemental Fig. S12**), suggesting that selective pressures from
470 antibiotics commonly present in the clinical environment enable *Lm* to acquire and retain more
471 ARGs. These findings suggest that *Lm* lineage I has a high risk of transmission from soil
472 environments to humans.
473
474 **DISCUSSION**
475
476 With in-depth pangenome analysis of soil-dwelling *Lm* that we isolated across the US, this study
477 reveals that bacterial pangenome fluidity is jointly shaped by environmental selection and
478 dispersal dynamics. We found that selective pressures from abiotic factors were mainly imposed
479 by climate (e.g., precipitation and temperature) and soil properties (e.g., aluminum, pH, and
480 molybdenum). Precipitation patterns likely influence bacterial pangenome fluidity by modulating
481 soil moisture, which in turn affects osmotic stress, salinity, and nutrient availability (Barnard et
482 al. 2013; Manzoni et al. 2012), potentially selecting for genes that enhance stress tolerance and

483 metabolic flexibility (Malik et al. 2024). Likewise, increased temperatures have been reported to
484 impact genetic variation by altering allele frequencies in genes associated with climate
485 adaptation, reducing genetic drift, and promoting homogeneous selection (De La Torre et al.
486 2019; Ning et al. 2020). Also, aluminum interacts with soil pH, where under acidic conditions (pH
487 < 5.5), it becomes solubilized into its bioavailable form, which has been found to yield selective
488 pressure on microbial populations (Ma et al. 2024; Zhang et al. 2022; Jones et al. 2019). In
489 addition, in our previous study (Liao et al. 2021), we identified molybdenum as the second most
490 important environmental factor influencing the nationwide distribution of *Listeria* in soils and
491 observed many of the top genes affected by positive selection were involved in the synthesis of
492 folate intermediate molybdopterin, a cofactor that binds molybdenum (Angelastro et al. 2017).
493 Consistent with these findings, we showed that molybdenum is a critical environmental stressor
494 that may influence gene gain and loss in *Lm* in this study.
495
496 While we found that abiotic factors tend to affect gene richness across multiple functions (e.g.,
497 carbohydrate, lipid, and inorganic ion transport and metabolism and transcription), the
498 pronounced effects were detected among genes involved in cell envelope synthesis and defense
499 mechanisms. In response to osmotic stress caused by precipitation fluctuations, Gram-positive
500 soil bacteria, such as *B. subtilis, Arthrobacter chlorophenolicus, Rhodococcus erythropolis,* and
501 *Mycobacterium pallens*, were found to modify their membrane composition and structure to
502 maintain cellular homeostasis (Bergholz et al. 2012; Karagulyan et al. 2022). Also, as the
503 temperature rises, bacterial membranes become more fluid, increasing permeability (Siliakus et
504 al. 2017), which can prompt homeoviscous adaptation by adjusting the fatty acid composition of
505 membrane lipids to counteract temperature-induced changes in fluidity (Siliakus et al. 2017;
506 Saunders et al. 2016; Bergholz et al. 2012). In addition, under low pH and high aluminum
507 conditions, bacteria can thicken their cell wall by upregulating peptidoglycan synthesis to
508 enhance its tolerance to aluminum toxicity, as observed in *R. erythropolis* (Ma et al. 2024).
509 Concurrent enrichment in defense-related genes may help bacteria adapt to other abiotic
510 stresses, such as oxidative stress and desiccation, by encoding stress response proteins crucial
511 for maintaining cellular function and viability under harsh conditions, as observed in *E. coli*
512 (Jaramillo-Riveri et al. 2022), *Acidithiobacillus ferrooxidans,* and *Leptospirillum ferriphilum*
513 (Claudia et al. 2024).
514
515 In addition to abiotic factors, we found that bacterial community composition also plays a role in
516 the pangenome fluidity of *Lm*. Bacterial phyla that showed the strongest associations were
517 Nitrospirae, Planctomycetes, Acidobacteria, and Cyanobacteria. Nitrospirae, commonly found in
518 oligotrophic environments (Martinez-Rabert et al. 2023; Zhao et al. 2023), are key players in
519 nitrification, particularly nitrite oxidation, which converts nitrite to nitrate (Daims et al. 2015).
520 Nitrate have been found to inhibit the growth of *Lm* (Majou and Christieans 2018; Hospital et al.
521 2017; Sullivan et al. 2012), which may select genetic machinery that involves retaining,
522 modulating, or losing genes in response to the adverse effects. Planctomycetes has been

observed co-existing with *Lm* in some environments, such as compost (e.g., soil amendments (Wang et al. 2022)) and spinach leaves (Culliney and Schmalenberger 2025), while a negative interaction between *Lm* and Acidobacteria has been reported (Liao et al. 2023; Falardeau et al. 2018; Locatelli et al. 2013). While the understanding of the interactions between Cyanobacteria and *Lm* is limited, it is known that Cyanobacteria could increase the production cyanotoxins (e.g., microcystins and anatoxin) to gain a competitive advantage under nitrogen-/phosphate-imbalance conditions or in the presence of co-occurring microbial species (Brandenburg et al. 2020; Tanvir et al. 2021; Holland and Kinnear 2013). Like abiotic factors, bacterial community composition was associated with genes across multiple functions (e.g., carbohydrate, lipid, and inorganic ion transport and metabolism and transcription). Particularly, the biotic factors appear to impose strong selective pressures on genes involved in cell envelope synthesis jointly with abiotic factors, and uniquely on genes involved in replication, recombination, and repair. Bacterial membranes often house signal transduction systems and histidine kinases crucial for communication, such as the Agr quorum-sensing system and LisRK two-component system (Vivant et al. 2014, 2015; Brunhede et al. 2020). These systems regulate the transcription or expression of genes associated with stress response, nutrient uptake, and membrane integrity, enabling bacteria to sense environmental cues and coordinate their behaviors within the community (Vivant et al. 2014, 2015; Brunhede et al. 2020). In addition, the association with genes involved in replication, recombination, and repair suggests that potential competitive interactions within the community may expose *Lm* to sources that cause DNA damage (Cornforth and Foster 2013; Pissaridou et al. 2018). In response, *Lm* may initiate DNA repair by activating its SOS system, which is regulated by repair genes such as *recA* (van der Veen et al. 2010).

Vast pangenome variation was observed among *Lm* evolutionary lineages, as a putative consequence of varying degrees of local environmental adaptation and dispersal dynamics. Lineage III was found to have a highly fluid pangenome but streamlined individual genome size with high GC content, which we attributed to strong adaptation to nutrient-limited conditions (e.g., carbon and nitrogen). Streamlined genomes enable efficient energy use while preserving the flexibility needed to thrive in challenging ecosystems, balancing adaptability with metabolic efficiency (Giovannoni et al. 2014; Ochman and Moran 2001). Consistent with our observation, a previous study reported that smaller genomes with higher GC content may lower the reproductive cost in carbon-limited soils in bacteria.(Chuckran et al. 2023) Environmental stresses have been shown to drive the loss of functionally redundant genes, resulting in genome streamlining, as observed in *Bradyrhizobium diazoefficiens* (Simonsen 2022). In response to varying abiotic and biotic selective pressures encountered in distinct soil conditions across the US, lineage III populations in different locations may lose different sets of genes, increasing the genome variation. In addition to differential local adaptation, strong dispersal limitation observed in this lineage could restrict gene flow across populations, enhancing its pangenome fluidity. In contrast, lineage I had a conserved pangenome, which was linked to frequent homogenizing

563 dispersal. Similar patterns have been observed in *Curtobacterium* and *Polynucleobacter*
564 *paneuropaeus*, where effective gene flow and high recombination efficacy facilitated by frequent
565 dispersal contribute to limited divergence and greater population coherence (Hoetzinger et al.
566 2021; Chase et al. 2019). With strong dispersal potential, *Lm* lineage I may spread from natural
567 environments like soils to nearby produce processing facilities through anthropogenic activities,
568 eventually contaminating food produce and infecting humans, evidenced by the epidemiological
569 links observed in this study and transmission routes documented previously (Liao et al. 2023;
570 Moura et al. 2016).
571
572 In summary, we provide critical insights into the ecological mechanisms underlying the bacterial
573 pangenome fluidity and suggest that environmental disturbance, especially shifts in climate
574 patterns, soil properties, and bacterial community composition, can profoundly affect the gene
575 gain and loss process in a bacterial species. Specifically, cell surface and defense mechanism-
576 related genes may be targeted by selective pressures from abiotic conditions, while genes
577 involved in replication, recombination, and repair may be uniquely influenced by bacterial
578 communities, in addition to cell surface-related genes jointly linked with abiotic factors. We also
579 highlight the important role of dispersal across geographic locations, along with environmental
580 selection, in shaping bacterial pangenome fluidity. Informed by the dispersal dynamics, we
581 additionally offer critical public health implications, emphasizing the need for targeted
582 environmental monitoring of *Listeria monocytogenes* lineage I to enhance source tracking of
583 foodborne outbreaks effectively.
584
585 **METHODS**
586
587 **Genomic data of soil *Lm* isolates, environmental data, and bacterial community data**
588
589 Genome assemblies of 177 *Lm* isolates obtained from soil samples collected by us from
590 minimally disturbed natural environments across the US and paired environmental data were
591 analyzed in this study. The genomic and environmental data were published in our previous
592 study investigating the adaptive evolution of *Listeria* (Liao et al. 2021). The methods for DNA
593 extraction (using QIAGEN QIAamp DNA MiniKit), whole genome sequencing (via Illumina
594 MiSeq, HiSeq, and NextSeq platforms), and sequencing data processing were detailed in Liao
595 et al. 2021. Genome assemblies of these isolates were of high quality, with < 300 contigs, an
596 N50 > 50,000, average coverage > 30X, consistent presence of the *sigB* allelic type (AT) in both
597 whole genome and PCR-based assays, no detected contamination using Kraken2 2.0.8 (Wood
598 and Salzberg 2014), and high genome completeness of 99.99% ± 0.01% (mean ± standard
599 deviation, SD) assessed by CheckM2 (Chklovski et al. 2023). Based on *sigB* allelic types (ATs)
600 and whole genome sequencing data, these *Lm* isolates were classified into lineage I (12
601 isolates), lineage II (39 isolates), and lineage III (126 isolates). Orthologous genes previously
602 identified using MMseqs2 (Steinegger and Söding 2017) and annotated for COG functions using
603 eggnog-mapper2 (Hernández-Plaza et al. 2023) were included in the data analysis in this study.

604 COGs abbreviations are as follows: C: Energy production and conversion; D: Cell cycle control,
605 cell division, chromosome partitioning; E: Amino acid transport and metabolism; F: Nucleotide
606 transport and metabolism; G: Carbohydrate transport and metabolism; H: Coenzyme transport
607 and metabolism; I: Lipid transport and metabolism; J: Translation, ribosomal structure, and
608 biogenesis; K: Transcription; L: Replication, recombination, and repair; M: Cell
609 wall/membrane/envelope biogenesis; N: Cell motility; O: Posttranslational modification, protein
610 turnover, chaperones; P: Inorganic ion transport and metabolism; Q: Secondary metabolites
611 biosynthesis, transport, and catabolism; T: Signal transduction mechanisms; U: Intracellular
612 trafficking, secretion, and vesicular transport; V: Defense mechanisms.
613
614 Environmental data analyzed in this study contained 34 variables, including three geolocation
615 variables (longitude, latitude, elevation), 17 soil physicochemical property variables (moisture,
616 total nitrogen [TN], total carbon [TC], pH, organic matter [OM], aluminum [Al], calcium [Ca],
617 copper [Cu], iron [Fe], potassium [K], magnesium [Mg], manganese [Mn], molybdenum [Mo],
618 sodium [Na], phosphorus [P], sulfur [S], and zinc [Zn]), four climate variables (precipitation, wind
619 speed, maximum and minimum temperatures), and 10 land use variables (proportions of open
620 water, barren, forest, shrubland, grassland, cropland, pasture, wetland, and developed [Dev]
621 open space with > 20% and < 20% impervious [IMP] cover in surrounding area). Detailed
622 methods for acquiring these environmental data were detailed in Liao et al. 2021.
623
624 Bacterial community data based on 16S rRNA gene amplicon sequencing were published in our
625 previous study comparing the genomic content of *Listeria* populations between soil and food-
626 associated environments (Liao et al. 2023). Briefly, total DNA was extracted from soil samples
627 using QIAGEN Power Soil Pro kits, and sequencing was performed on a MiSeq platform with 2
628 × 250 bp paired-end reads. Raw sequence reads were processed using QIIME2 to remove
629 noises, identify OTUs, classify taxonomy, and calculate beta diversity, including weighted
630 UniFrac distances. Bacterial phyla were referred to as biotic factors in this study.
631
632 **Gene richness analysis**
633
634 VPA was conducted to assess the contributions of **i)** geolocation, soil properties, climate, and
635 surrounding land use, and **ii)** abiotic factors (i.e., all four groups in **i)** combined) and biotic factors
636 (i.e., the relative abundance of bacterial phyla) to the variation of gene richness overall and for
637 COGs at both species and lineage levels. VPA was computed and the adjusted R^2 values were
638 visualized in a Venn diagram using the 'varpart' function in the vegan package 2.6-4 in R. To
639 assess the significance of the explained variation, a permutation test was performed, in which
640 gene richness was randomly shuffled across genomes 100 times to generate a null distribution
641 for VPA and the resulting explained variation (i.e., expected explained variation) was compared
642 against the observed explained variation using a one-sided test, with significance set at $P < 0.05$.
643 Spearman's correlation was further conducted to evaluate associations between individual

644 abiotic and biotic factors and gene richness overall and for COGs. To account for multiple testing,
645 *P*-values were adjusted using the Benjamini-Hochberg false discovery rate (BH-FDR) method,
646 with adjusted $P < 0.05$ considered significant.
647
648 **RDA, machine learning models, and identification of abiotic/biotic-linked genes**
649
650 To select between RDA and canonical correspondence analysis (CCA) for exploring
651 relationships between abiotic/biotic variables and accessory genomes, we performed detrended
652 correspondence analysis (DCA) (Lepš and Šmilauer 2003). The first DCA axis length was 0.025,
653 shorter than three standard deviations, indicating that the data is relatively homogeneous and
654 that linear methods, like RDA, are more appropriate for our dataset (Lepš and Šmilauer 2003).
655 Prior to conducting RDA, abiotic factors were standardized to account for the different units of
656 measurement. L1 regularization (Lasso) was applied for feature selection to reduce
657 multicollinearity and enhance model interpretability (Hastie et al. 2009). The Lasso procedure
658 selected 31 out of 34 abiotic factors and 17 out of 27 bacterial phyla important for the *Lm*
659 accessory genome, which were included in RDA. The length and direction of arrows in the RDA
660 biplot indicate the relative importance and correlation of each abiotic/biotic factor with the
661 ordination axes. DCA and RDA were performed using the 'decorana' function in the vegan
662 package 2.6-4 in R and the 'skbio.stats.ordination.rda' function in the skbio library in Python
663 3.6.8, respectively.
664
665 To predict the RDA ordination axes from abiotic and biotic factors separately, we implemented
666 a machine learning-based framework using Python 3.6.8 reported in our previous study (Goh et
667 al. 2024). In brief, the data were preprocessed and split into training (80%) and testing (20%)
668 sets. The training set was further partitioned into five folds for cross-validation, during which
669 various regression models were trained and evaluated with optimized hyperparameters (see
670 **Supplemental Table S5**). To identify the best-performing model, we conducted a random
671 search across 100 parameter configurations for each algorithm, including regressors based on
672 random forests, neural networks, gradient boosting, support vector machines, and *k*-nearest
673 neighbors. The model with the highest R^2 score was then fine-tuned through exhaustive
674 hyperparameter exploration to achieve optimal performance. The importance of the features to
675 the prediction was quantified using SHAP (Lundberg and Lee 2017).
676
677 Two-sided Mann-Whitney *U* tests followed by BH-FDR adjustment were employed to compare
678 abiotic and biotic factors between genomes with and without a given accessory gene. Genes
679 with an adjusted $P < 0.05$ were considered significantly linked to a particular abiotic factor
680 (defined as "abiotic-linked genes") or linked to a particular phylum (defined as "biotic-linked
681 genes"). Functional enrichment analysis was conducted using a binomial distribution model
682 described in Liao et al. 2023 to identify COGs significantly enriched among abiotic-linked genes,
683 biotic-linked genes, and genes exclusively linked to abiotic or biotic factors.

684
685 **Abiotic and biotic environmental conditions of *Lm* lineages and distance-decay**
686 **relationship**
687
688 We re-visualized the previously published spatial distribution map of 177 *Lm* genomes
689 (**Extended Data Fig. 5** in Liao et al. 2021) using the Mercator Projection and the Basemap
690 Matplotlib Toolkit 1.2.1 in Python 3.6.8. Kruskal-Wallis tests followed by BH-FDR correction were
691 performed to identify significant differences in abiotic and biotic factors among *Lm* lineages. For
692 factors with an adjusted Kruskal-Wallis $P < 0.05$, pairwise differences between two lineages
693 were further assessed using two-sided Mann-Whitney *U* tests followed by BH-FDR correction.
694 Additionally, MDS analysis was performed using Euclidean distance for abiotic factors, and
695 weighted UniFrac distance for OTUs to compare abiotic conditions and bacterial community
696 compositions among *Lm* lineages followed by a PERMANOVA test to assess clustering
697 significance, with $P < 0.05$ indicating significant clustering by lineages. The distance–decay
698 relationship was inferred by linear regression between geographical distances (km) and
699 biological similarities (indicated by ANI) between genomes for each lineage. A steeper negative
700 slope of the fitted regression line with a higher R^2 indicates a stronger distance-decay
701 relationship.
702
703 **Pangenome features of *Lm* lineages**
704
705 To predict the pangenome size given 100 genomes for each *Lm* lineage, we applied the power
706 law function cN^{γ}, calculated in **Extended Data Fig. 7** in Liao et al. 2021. The specific functions
707 used were $n_{pan} = 2723N^{0.093}$ for lineage I, $n_{pan} = 2681N^{0.122}$ for lineage II, and $n_{pan} = 2490N^{0.159}$
708 for lineage III, where *N* is the number of genomes (Liao et al. 2021). Fisher's exact tests followed
709 by BH-FDR correction were performed in Python 3.6.8 to identify genes having significantly
710 different frequencies among lineages, referred to as "lineage-associated genes" (adjusted $P <$
711 0.05). Functional enrichment analysis was performed using a binomial distribution model
712 described in Liao et al. 2023 to identify COGs significantly enriched among lineage-associated
713 genes.
714
715 Additionally, various genomic elements were analyzed, including genome size and GC content
716 (previously reported in Liao et al. 2021), virulence factors (LIPI-1, -3, -4 and *inl* genes), SSI-1
717 and 2 (previously reported in Liao et al. 2023), as well as ARGs and MGEs, including IS,
718 transposons, prophages, and plasmids (previously reported in Goh et al. 2024). In brief,
719 virulence factors, SSI, and ARG were previously identified through BLASTN searches against
720 reference databases sourced from the BIGSdb-*Lm* platform (Moura et al. 2016). Reference
721 virulence factors include LIPI-1 (*prfA, plcA, hly, mpl, actA, plcB*), LIPI-3 (*llsAGHXBYDP*), LIPI-4
722 (*LM9005581_70009 - LM9005581_70014*), and *inl* genes (*inlAB, C, E, F, G, H, J, K, I, P*).
723 Reference SSI-1 include *lmo0444 - lmo0448* and SSI-2 include *lin0464 - lin0465*. Reference

ARGs includes *aacA4, aadC, aadE, aphA, cat_CHL, dfrD, dfrK, ermB, ermG, fexA, fosX, lin, mprF, lnuA, lnuG, mefA, mphB, msrD, norB, penA, qnrB, str, sul, tetM,* and *tetS*. ISEScan (Xie and Tang 2017), TnFinder (Ross et al. 2021), PHASTEST (Wishart et al. 2023), and PlasmidFinder (Carattoli et al. 2014) were previously used for the prediction of IS, transposons, prophages, and plasmids, respectively. While these genomic elements were previously reported, they were not compared among *Lm* lineages. Here, Kruskal-Wallis tests were used to assess the differences in genome size, GC content, and MGEs among lineages, followed by two-sided Mann-Whitney *U* tests with BH-FDR correction to assess the pairwise differences. Fisher's exact tests with BH-FDR correction were used to compare the frequency of virulence factors, SSI, and ARGs among lineages. An adjusted $P < 0.05$ was considered statistically significant.

Identification of epidemiological links between *Lm* soil and clinical isolates

The NCBI Pathogen Detection Isolates Browser (National Center for Biotechnology Information 2016) was used in 2019 to identify human clinical *Lm* isolates found within the same single-linkage SNP cluster as with one or more of the 177 soil *Lm* isolates included in this study. Single-linkage SNP cluster is defined as a cluster of isolates differing by no more than 50 SNPs (Dallman et al. 2018). A total of 186 closely related clinical isolates were identified, all of which met the genome assembly quality criteria outlined in Liao et al. 2021 (i.e., < 300 contigs, N50 > 50,000), with a high genome completeness of 99.98% ± 0.01% (mean ± SD, **Supplemental Table S6**). A phylogenetic tree including both soil and closely related clinical isolates was constructed using RAxML-8.2.13 (Stamatakis 2014) based core SNPs identified using kSNP4 (Hall and Nisbet 2023). The GTR + G (General Time Reversible + gamma distribution) model selected by jModelTest (Stamatakis 2014), along with ascertainment bias correction and 1,000 bootstrap repetitions, was applied in the tree construction. The phylogenetic tree was visualized using the Interactive Tree of Life (iTOL) 6 webserver (Letunic and Bork 2024).

To identify epidemiologic links between soil and clinical *Lm* isolates, cgMLST was profiled, in which ATs for cgMLST genes were assigned based on the database available on the BIGSdb-*Lm* platform (Moura et al. 2016). Isolates differing by up to seven allelic mismatches in cgMLST profiles were considered epidemiologically linked (Moura et al. 2016). Genome size, GC content, and total counts for virulence factors, SSI-1 and 2, ARGs, and normalized abundance of MGEs (i.e., IS, transposons, and prophages) were compared between soil isolate L7-1173 and its epidemiologically linked clinical isolates (15 total) using the Wilcoxon rank-sum test, with $P < 0.05$ considered statistically significant. The other two soil isolates (L7-0745 and L7-0846) were epidemiologically linked to one clinical isolate each, so they were not included in the statistical tests.

DATA ACCESS

The whole genome sequencing data for 177 *Listeria monocytogenes* soil isolates used in this study have been submitted to the NCBI BioProject database (https://www.ncbi.nlm.nih.gov/bioproject/) under accession numbers PRJNA514286 and PRJNA561882, along with associated environmental data, which were previously published in (Liao et al. 2021). The 16S rRNA gene amplicon sequencing data used in this study have been submitted to the NCBI BioProject database (https://www.ncbi.nlm.nih.gov/bioproject/) under accession number PRJNA749132 and were published in (Liao et al. 2023) NCBI GenBank accession numbers of the genomic assemblies for clinical Lm isolates identified via the NCBI Pathogen Detection Isolates Browser are provided in **Supplemental Table S6**. All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. Code to replicate all analyses is publicly available at https://github.com/leaph-lab/Lm_pangenome_MS.

COMPETING INTEREST STATEMENT

The authors declare no competing interests.

ACKNOWLEDGMENTS

We are grateful to the members of LEAPH (the Liao laboratory) for their enriching discussions. This work was funded by the 4-VA (J.L.) and the Virginia Tech Center for Emerging, Zoonotic, and Arthropod-borne Pathogens Interdisciplinary Graduate Education Program in Infectious Disease (Y.-X.G.).

AUTHOR CONTRIBUTIONS

Conceptualization, J.L.; Methodology, Y.-X.G., F.H., H.Z., J.L.; Formal Analysis, Y.-X.G., F.H., H.Z., J.L.; Writing—original draft, Y.-X.G.; Writing—review & editing, J.L.; Visualization, Y.-X.G., F.H., H.Z., J.L.; Supervision, J.L.; Funding Acquisition, J.L.

REFERENCES

Albalat R, Cañestro C. 2016. Evolution by gene loss. *Nat Rev Genet* **17**: 379–391. doi: 10.1038/nrg.2016.39

Angelastro A, Dawson WM, Luk LYP, Loveridge EJ, Allemann RK. 2017. Chemoenzymatic assembly of isotopically labeled folates. *J Am Chem Soc* **139**: 13047–13054. doi: 10.1021/jacs.7b06358

Barnard RL, Osborne CA, Firestone MK. 2013. Responses of soil bacterial and fungal communities to extreme desiccation and rewetting. *ISME J* **7**: 2229–2241. doi: 10.1038/ismej.2013.104

Behari J, Youngman P. 1998. Regulation of *hly* expression in *Listeria monocytogenes* by carbon sources and pH occurs through separate mechanisms mediated by PrfA. *Infect Immun* **66**: 3635–3642. doi: 10.1128/IAI.66.8.3635-3642.1998

803 Bell T. 2010. Experimental tests of the bacterial distance–decay relationship. *ISME J* **4**: 1357–
804 1365. doi: 10.1038/ismej.2010.77
805 Bergholz TM, Bowen B, Wiedmann M, Boor KJ. 2012. *Listeria monocytogenes* shows
806 temperature-dependent and -independent responses to salt stress, including responses
807 that induce cross-protection against other stresses. *Appl Environ Microbiol* **78**: 2602–2612.
808 doi: 10.1128/AEM.07658-11
809 Bierne H, Sabet C, Personnic N, Cossart P. 2007. Internalins: a complex family of leucine-rich
810 repeat-containing proteins in *Listeria monocytogenes*. *Microbes Infect* **9**: 1156–1166. doi:
811 10.1016/j.micinf.2007.05.003
812 Bourne EC, Bocedi G, Travis JMJ, Pakeman RJ, Brooker RW, Schiffers K. 2014. Between
813 migration load and evolutionary rescue: dispersal, adaptation and the response of spatially
814 structured populations to environmental change. *Proc Biol Sci* **281**: 20132795. doi:
815 10.1098/rspb.2013.2795
816 Brandenburg K, Siebers L, Keuskamp J, Jephcott TG, Van de Waal DB. 2020. Effects of nutrient
817 limitation on the synthesis of N-rich phytoplankton toxins: a meta-analysis. *Toxins (Basel)*
818 **12**: 221. doi: 10.3390/toxins12040221
819 Brockhurst MA, Harrison E, Hall JPJ, Richards T, McNally A, MacLean C. 2019. The ecology
820 and evolution of pangenomes. *Curr Biol* **29**: R1094–R1103. doi: 10.1016/j.cub.2019.08.012
821 Brunhede MZ, Santos PT Dos, Gal L, Garmyn D, Kallipolitis BH, Piveteau P. 2020. LisRK is
822 required for optimal fitness of *Listeria monocytogenes* in soil. *FEMS Microbiol Lett* **367**:
823 fnaa188. doi: 10.1093/femsle/fnaa188
824 Carattoli A, Zankari E, García-Fernández A, Voldby Larsen M, Lund O, Villa L, Møller Aarestrup
825 F, Hasman H. 2014. *In silico* detection and typing of plasmids using PlasmidFinder and
826 plasmid multilocus sequence typing. *Antimicrob Agents Chemother* **58**: 3895–3903. doi:
827 10.1128/AAC.02412-14
828 Chase AB, Arevalo P, Brodie EL, Polz MF, Karaoz U, Martiny JBH. 2019. Maintenance of
829 sympatric and allopatric populations in free-living terrestrial bacteria. *mBio* **10**:e02361-19.
830 doi: 10.1128/mBio.02361-19
831 Chklovski A, Parks DH, Woodcroft BJ, Tyson GW. 2023. CheckM2: a rapid, scalable and
832 accurate tool for assessing microbial genome quality using machine learning. *Nat Methods*
833 **20**: 1203–1212. doi: 10.1038/s41592-023-01940-w
834 Chuckran PF, Flagg C, Propster J, Rutherford WA, Sieradzki ET, Blazewicz SJ, Hungate B, Pett-
835 Ridge J, Schwartz E, Dijkstra P. 2023. Edaphic controls on genome size and GC content of
836 bacteria in soil microbial communities. *Soil Biol Biochem* **178**: 108935. doi:
837 10.1016/j.soilbio.2022.108935
838 Claudia M-V, Javiera A-A, Sebastián N-S, José F-R, Gloria L. 2024. Interplay between
839 desiccation and oxidative stress responses in iron-oxidizing acidophilic bacteria. *J*
840 *Biotechnol* **383**: 64–72. doi: 10.1016/j.jbiotec.2024.01.017

841 Conrad RE, Viver T, Gago JF, Hatt JK, Venter SN, Rossello-Mora R, Konstantinidis KT. 2022.
842 Toward quantifying the adaptive role of bacterial pangenomes during environmental
843 perturbations. *ISME J* **16**: 1222–1234. doi: 10.1038/s41396-021-01149-9
844 Cornforth DM, Foster KR. 2013. Competition sensing: the social side of bacterial stress
845 responses. *Nat Rev Microbiol* **11**: 285–293. doi: 10.1038/nrmicro2977
846 Cotter PD, Draper LA, Lawton EM, Daly KM, Groeger DS, Casey PG, Ross RP, Hill C. 2008.
847 Listeriolysin S, a novel peptide haemolysin associated with a subset of lineage I *Listeria*
848 *monocytogenes*. *PLoS Pathog* **4**: e1000144. doi: 10.1371/journal.ppat.1000144
849 Crits-Christoph A, Olm MR, Diamond S, Bouma-Gregson K, Banfield JF. 2020. Soil bacterial
850 populations are shaped by recombination and gene-specific selection across a grassland
851 meadow. *ISME J* **14**: 1834–1846. doi: 10.1038/s41396-020-0655-x
852 Culliney P, Schmalenberger A. 2025. Bacterial community structure analysis on *Listeria*
853 *monocytogenes* inoculated spinach leaves is affected by PCR based methods to exclude
854 chloroplast co-amplification. *The Microbe* **6**: 100258. doi: 10.1016/j.microb.2025.100258
855 Cummins EA, Hall RJ, McInerney JO, McNally A. 2022. Prokaryote pangenomes are dynamic
856 entities. *Curr Opin Microbiol* **66**: 73–78. doi: 10.1016/j.mib.2022.01.005
857 Daims H, Lebedeva E V., Pjevac P, Han P, Herbold C, Albertsen M, Jehmlich N, Palatinszky M,
858 Vierheilig J, Bulaev A, et al. 2015. Complete nitrification by *Nitrospira* bacteria. *Nature* **528**:
859 504–509. doi: 10.1038/nature16461
860 Dallman T, Ashton P, Schafer U, Jironkin A, Painset A, Shaaban S, Hartman H, Myers R,
861 Underwood A, Jenkins C, et al. 2018. SnapperDB: a database solution for routine
862 sequencing analysis of bacterial isolates. *Bioinformatics* **34**: 3028–3029. doi:
863 10.1093/bioinformatics/bty212
864 De La Torre AR, Wilhite B, Neale DB. 2019. Environmental genome-wide association reveals
865 climate adaptation is shaped by subtle to moderate allele frequency shifts in loblolly pine.
866 *Genome Biol Evol* **11**: 2976–2989. doi: 10.1093/gbe/evz220
867 den Bakker HC, Bowen BM, Rodriguez-Rivera LD, Wiedmann M. 2012. FSL J1-208, a virulent
868 uncommon phylogenetic lineage IV *Listeria monocytogenes* strain with a small chromosome
869 size and a putative virulence plasmid carrying internalin-like genes. *Appl Environ Microbiol*
870 **78**: 1876–1889. doi: 10.1128/AEM.06969-11
871 den Bakker HC, Desjardins CA, Griggs AD, Peters JE, Zeng Q, Young SK, Kodira CD, Yandava
872 C, Hepburn TA, Haas BJ, et al. 2013. Evolutionary dynamics of the accessory genome of
873 *Listeria monocytogenes*. *PLoS One* **8**: e67511. doi: 10.1371/journal.pone.0067511
874 Falardeau J, Walji K, Haure M, Fong K, Taylor G, Ma Y, Smukler S, Wang S. 2018. Native
875 bacterial communities and *Listeria monocytogenes* survival in soils collected from the Lower
876 Mainland of British Columbia, Canada. *Can J Microbiol* **64**: 695–705. doi: 10.1139/cjm-
877 2018-0115
878 Freitag NE, Port GC, Miner MD. 2009. *Listeria monocytogenes* — from saprophyte to
879 intracellular pathogen. *Nat Rev Microbiol* **7**: 623–628. doi: 10.1038/nrmicro2171

880 Garant D, Forde SE, Hendry AP. 2007. The multifarious effects of dispersal and gene flow on
881 contemporary adaptation. *Funct Ecol* **21**: 434–443. doi: 10.1111/j.1365-2435.2006.01228.x
882 Giovannoni SJ, Cameron Thrash J, Temperton B. 2014. Implications of streamlining theory for
883 microbial ecology. *ISME J* **8**: 1553–1565. doi: 10.1038/ismej.2014.60
884 Goh Y-X, Anupoju SMB, Nguyen A, Zhang H, Ponder M, Krometis L-A, Pruden A, Liao J. 2024.
885 Evidence of horizontal gene transfer and environmental selection impacting antibiotic
886 resistance evolution in soil-dwelling *Listeria*. *Nat Commun* **15**: 10034. doi: 10.1038/s41467-
887 024-54459-9
888 Hall BG, Nisbet J. 2023. Building phylogenetic trees from genome sequences with kSNP4. *Mol*
889 *Biol Evol* **40**: msad235. doi: 10.1093/molbev/msad235
890 Harter E, Wagner EM, Zaiser A, Halecker S, Wagner M, Rychli K. 2017. Stress survival islet 2,
891 predominantly present in *Listeria monocytogenes* strains of sequence type 121, is involved
892 in the alkaline and oxidative stress responses. *Appl Environ Microbiol* **83**: e00827-17. doi:
893 10.1128/AEM.00827-17
894 Hastie T, Tibshirani R, Friedman J. 2009. *The elements of statistical learning*. Springer, New
895 York, NY.
896 Hernández-Plaza A, Szklarczyk D, Botas J, Cantalapiedra CP, Giner-Lamia J, Mende DR,
897 Kirsch R, Rattei T, Letunic I, Jensen LJ, et al. 2023. eggNOG 6.0: enabling comparative
898 genomics across 12 535 organisms. *Nucleic Acids Res* **51**: D389–D394. doi:
899 10.1093/nar/gkac1022
900 Hoetzinger M, Pitt A, Huemer A, Hahn MW. 2021. Continental-scale gene flow prevents
901 allopatric divergence of pelagic freshwater bacteria. *Genome Biol Evol* **13**: evab019. doi:
902 10.1093/gbe/evab019
903 Hoffmann S, Batz MB, Morris JG. 2012. Annual cost of illness and quality-adjusted life year
904 losses in the United States due to 14 foodborne pathogens. *J Food Prot* **75**: 1292–1302.
905 doi: 10.4315/0362-028X.JFP-11-417
906 Holland A, Kinnear S. 2013. Interpreting the possible ecological role(s) of cyanotoxins:
907 compounds for competitive advantage and/or physiological aide? *Mar Drugs* **11**: 2239–
908 2258. doi: 10.3390/md11072239
909 Hospital XF, Hierro E, Arnau J, Carballo J, Aguirre JS, Gratacós-Cubarsí M, Fernández M. 2017.
910 Effect of nitrate and nitrite on *Listeria* and selected spoilage bacteria inoculated in dry-cured
911 ham. *Food Res Int* **101**: 82–87. doi: 10.1016/j.foodres.2017.08.039
912 Jaramillo-Riveri S, Broughton J, McVey A, Pilizota T, Scott M, El Karoui M. 2022. Growth-
913 dependent heterogeneity in the DNA damage response in *Escherichia coli*. *Mol Syst Biol*
914 **18**: e10441. doi: 10.15252/msb.202110441
915 Jones DL, Cooledge EC, Hoyle FC, Griffiths RI, Murphy D V. 2019. pH and exchangeable
916 aluminum are major regulators of microbial energy flow and carbon use efficiency in soil
917 microbial communities. *Soil Biol Biochem* **138**: 107584. doi: 10.1016/j.soilbio.2019.107584

Karagulyan M, Goebel M-O, Diehl D, Abu Quba AA, Kästner M, Bachmann J, Wick LY, Schaumann GE, Miltner A. 2022. Water stress-driven changes in bacterial cell surface properties. *Appl Environ Microbiol* **88**: e0073222. doi: 10.1128/aem.00732-22

Koopmans MM, Brouwer MC, Vázquez-Boland JA, van de Beek D. 2023. Human listeriosis. *Clin Microbiol Rev* **36**: e0006019. doi: 10.1128/cmr.00060-19

Legendre P, Borcard D, Peres-Neto PR. 2005. Analyzing beta diversity: partitioning the spatial variation of community composition data. *Ecol Monogr* **75**: 435–450. doi: 10.1890/05-0549

Lepš J, Šmilauer P. 2003. *Multivariate analysis of ecological data using CANOCO*. Cambridge University Press, Cambridge, England.

Letunic I, Bork P. 2024. Interactive Tree of Life (iTOL) v6: recent updates to the phylogenetic tree display and annotation tool. *Nucleic Acids Res* **52**: W78–W82. doi: 10.1093/nar/gkae268

Liao J, Guo X, Li S, Anupoju SMB, Cheng RA, Weller DL, Sullivan G, Zhang H, Deng X, Wiedmann M. 2023. Comparative genomics unveils extensive genomic variation between populations of *Listeria* species in natural and food-associated environments. *ISME Commun* **3**: 85. doi: 10.1038/s43705-023-00293-x

Liao J, Guo X, Weller DL, Pollak S, Buckley DH, Wiedmann M, Cordero OX. 2021. Nationwide genomic atlas of soil-dwelling *Listeria* reveals effects of selection and population ecology on pangenome evolution. *Nat Microbiol* **6**: 1021–1030. doi: 10.1038/s41564-021-00935-7

Linke K, Rückerl I, Brugger K, Karpiskova R, Walland J, Muri-Klinger S, Tichy A, Wagner M, Stessl B. 2014. Reservoirs of *Listeria* species in three environmental ecosystems. *Appl Environ Microbiol* **80**: 5583–5592. doi: 10.1128/AEM.01018-14

Locatelli A, Spor A, Jolivet C, Piveteau P, Hartmann A. 2013. Biotic and abiotic soil properties influence survival of *Listeria monocytogenes* in soil. *PLoS One* **8**: e75969. doi: 10.1371/journal.pone.0075969

Lundberg SM, Lee SI. 2017. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, vols 2017-December.

Ma Z, Jiang M, Liu C, Wang E, Bai Y, Yuan MM, Shi S, Zhou J, Ding J, Xie Y, et al. 2024. Quinolone-mediated metabolic cross-feeding develops aluminium tolerance in soil microbial consortia. *Nat Commun* **15**: 10148. doi: 10.1038/s41467-024-54616-0

Majou D, Christieans S. 2018. Mechanisms of the bactericidal effects of nitrate and nitrite in cured meats. *Meat Sci* **145**: 273–284. doi: 10.1016/j.meatsci.2018.06.013

Malik AA, Martiny JBH, Ribeiro A, Sheridan PO, Weihe C, Brodie EL, Allison SD. 2024. Bacterial population-level trade-offs between drought tolerance and resource acquisition traits impact decomposition. *ISME J* **18**: wrae224. doi: 10.1093/ismejo/wrae224

Manzoni S, Schimel JP, Porporato A. 2012. Responses of soil microbial communities to water stress: results from a meta-analysis. *Ecology* **93**: 930–938. doi: 10.1890/11-0026.1

Martinez-Rabert E, Smith CJ, Sloan WT, Gonzalez-Cabaleiro R. 2023. Competitive and substrate limited environments drive metabolic heterogeneity for comammox *Nitrospira*. *ISME Commun* **3**: 91. doi: 10.1038/s43705-023-00288-8

958 Martinson JN V, Chacón JM, Smith BA, Villarreal AR, Hunter RC, Harcombe WR. 2023.
959 Mutualism reduces the severity of gene disruptions in predictable ways across microbial
960 communities. *ISME J* **17**: 2270–2278. doi: 10.1038/s41396-023-01534-6
961 Maury MM, Bracq-Dieye H, Huang L, Vales G, Lavina M, Thouvenot P, Disson O, Leclercq A,
962 Brisse S, Lecuit M. 2019. Hypervirulent *Listeria monocytogenes* clones' adaption to
963 mammalian gut accounts for their association with dairy products. *Nat Commun* **10**: 2488.
964 doi: 10.1038/s41467-019-10380-0
965 Maury MM, Tsai Y-H, Charlier C, Touchon M, Chenal-Francisque V, Leclercq A, Criscuolo A,
966 Gaultier C, Roussel S, Brisabois A, et al. 2016. Uncovering *Listeria monocytogenes*
967 hypervirulence by harnessing its biodiversity. *Nat Genet* **48**: 308–313. doi: 10.1038/ng.3501
968 McInerney JO, McNally A, O'Connell MJ. 2017. Why prokaryotes have pangenomes. *Nat*
969 *Microbiol* **2**: 17040. doi: 10.1038/nmicrobiol.2017.40
970 Moura A, Criscuolo A, Pouseele H, Maury MM, Leclercq A, Tarr C, Björkman JT, Dallman T,
971 Reimer A, Enouf V, et al. 2016. Whole genome-based population biology and
972 epidemiological surveillance of *Listeria monocytogenes*. *Nat Microbiol* **2**: 16185. doi:
973 10.1038/nmicrobiol.2016.185
974 National Center for Biotechnology Information. 2016. NCBI pathogen detection isolates browser.
975 https://www.ncbi.nlm.nih.gov/pathogens/isolates (Accessed December 30, 2019).
976 Ning D, Yuan M, Wu L, Zhang Y, Guo X, Zhou X, Yang Y, Arkin AP, Firestone MK, Zhou J. 2020.
977 A quantitative framework reveals ecological drivers of grassland microbial community
978 assembly in response to warming. *Nat Commun* **11**: 4717. doi: 10.1038/s41467-020-18560-
979 z
980 Nowell RW, Green S, Laue BE, Sharp PM. 2014. The extent of genome flux and its role in the
981 differentiation of bacterial lineages. *Genome Biol Evol* **6**: 1514–1529. doi:
982 10.1093/gbe/evu123
983 Ochman H, Moran NA. 2001. Genes lost and genes found: evolution of bacterial pathogenesis
984 and symbiosis. *Science* **292**: 1096–1099. doi: 10.1126/science.1058543
985 Orsi RH, Bakker HC den, Wiedmann M. 2011. *Listeria monocytogenes* lineages: genomics,
986 evolution, ecology, and phenotypic characteristics. *Int J Med Microbiol* **301**: 79–96. doi:
987 10.1016/j.ijmm.2010.05.002
988 Orsi RH, Wiedmann M. 2016. Characteristics and distribution of *Listeria* spp., including *Listeria*
989 species newly described since 2009. *Appl Microbiol Biotechnol* **100**: 5273–5287. doi:
990 10.1007/s00253-016-7552-2
991 Pissaridou P, Allsopp LP, Wettstadt S, Howard SA, Mavridou DAI, Filloux A. 2018. The
992 *Pseudomonas aeruginosa* T6SS-VgrG1b spike is topped by a PAAR protein eliciting DNA
993 damage to bacterial competitors. *Proc Natl Acad Sci USA* **115**: 12519–12524. doi:
994 10.1073/pnas.1814181115
995 Richardson JL, Urban MC, Bolnick DI, Skelly DK. 2014. Microgeographic adaptation and the
996 spatial scale of evolution. *Trends Ecol Evol* **29**: 165–176. doi: 10.1016/j.tree.2014.01.002

997 Rodriguez-Valera F, Martin-Cuadrado A-B, Rodriguez-Brito B, Pašić L, Thingstad TF, Rohwer
998 F, Mira A. 2009. Explaining microbial population genomics through phage predation. *Nat*
999 *Rev Microbiol* **7**: 828–836. doi: 10.1038/nrmicro2235
1000 Ross K, Varani AM, Snesrud E, Huang H, Alvarenga DO, Zhang J, Wu C, McGann P, Chandler
1001 M. 2021. TnCentral: a prokaryotic transposable element database and web portal for
1002 transposon analysis. *mBio* **12**: e0206021. doi: 10.1128/mBio.02060-21
1003 Rouli L, Merhej V, Fournier P-E, Raoult D. 2015. The bacterial pangenome as a new tool for
1004 analysing pathogenic bacteria. *New Microbes New Infect* **7**: 72–85. doi:
1005 10.1016/j.nmni.2015.06.005
1006 Ryan S, Begley M, Hill C, Gahan CGM. 2010. A five-gene stress survival islet (SSI-1) that
1007 contributes to the growth of *Listeria monocytogenes* in suboptimal conditions. *J Appl*
1008 *Microbiol* **109**: 984–995. doi: 10.1111/j.1365-2672.2010.04726.x
1009 Saunders LP, Sen S, Wilkinson BJ, Gatto C. 2016. Insights into the mechanism of homeoviscous
1010 adaptation to low temperature in branched-chain fatty acid-containing bacteria through
1011 modeling FabH kinetics from the foodborne pathogen *Listeria monocytogenes*. *Front*
1012 *Microbiol* **7**: 1386. doi: 10.3389/fmicb.2016.01386
1013 Siliakus MF, van der Oost J, Kengen SWM. 2017. Adaptations of archaeal and bacterial
1014 membranes to variations in temperature, pH and pressure. *Extremophiles* **21**: 651–670. doi:
1015 10.1007/s00792-017-0939-x
1016 Simonsen AK. 2022. Environmental stress leads to genome streamlining in a widely distributed
1017 species of soil bacteria. *ISME J* 16: 423–434. doi: 10.1038/s41396-021-01082-x
1018 Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large
1019 phylogenies. *Bioinformatics* **30**: 1312–1313. doi: 10.1093/bioinformatics/btu033
1020 Steinegger M, Söding J. 2017. MMseqs2 enables sensitive protein sequence searching for the
1021 analysis of massive data sets. *Nat Biotechnol* **35**: 1026–1028. doi: 10.1038/nbt.3988
1022 Sullivan GA, Jackson-Davis AL, Niebuhr SE, Xi Y, Schrader KD, Sebranek JG, Dickson JS.
1023 2012. Inhibition of *Listeria monocytogenes* using natural antimicrobials in no-nitrate-or-
1024 nitrite-added ham. *J Food Prot* **75**: 1071–1076. doi: 10.4315/0362-028X.JFP-11-511
1025 Tanvir RU, Hu Z, Zhang Y, Lu J. 2021. Cyanobacterial community succession and associated
1026 cyanotoxin production in hypereutrophic and eutrophic freshwaters. *Environmental Pollution*
1027 **290**: 118056. doi: 10.1016/j.envpol.2021.118056
1028 Thomas CM, Nielsen KM. 2005. Mechanisms of, and barriers to, horizontal gene transfer
1029 between bacteria. *Nat Rev Microbiol* 3: 711–721. doi: 10.1038/nrmicro1234
1030 US Food and Drug Administration. 2020. Get the facts about *Listeria*.
1031 https://www.fda.gov/animal-veterinary/animal-health-literacy/get-facts-about-listeria.
1032 van der Veen S, van Schalkwijk S, Molenaar D, de Vos WM, Abee T, Wells-Bennik MHJ. 2010.
1033 The SOS response of *Listeria monocytogenes* is involved in stress resistance and
1034 mutagenesis. *Microbiology* **156**: 374–384. doi: 10.1099/mic.0.035196-0

1035 Vivant A-L, Garmyn D, Gal L, Hartmann A, Piveteau P. 2015. Survival of *Listeria monocytogenes*
1036 in soil requires AgrA-mediated regulation. *Appl Environ Microbiol* **81**: 5073–5084. doi:
1037 10.1128/AEM.04134-14
1038 Vivant A-L, Garmyn D, Gal L, Piveteau P. 2014. The Agr communication system provides a
1039 benefit to the populations of *Listeria monocytogenes* in soil. *Front Cell Infect Microbiol* **4**:
1040 160. doi: 10.3389/fcimb.2014.00160
1041 Vogwill T, Fenton A, Brockhurst MA. 2010. How does spatial dispersal network affect the
1042 evolution of parasite local adaptation? *Evolution* **64**: 1795–1801. doi: 10.1111/j.1558-
1043 5646.2009.00937.x
1044 Wang C, Yu Q-Y, Ji N-N, Zheng Y, Taylor JW, Guo L-D, Gao C. 2023. Bacterial genome size
1045 and gene functional diversity negatively correlate with taxonomic diversity along a pH
1046 gradient. *Nat Commun* **14**: 7437. doi: 10.1038/s41467-023-43297-w
1047 Wang H, Shankar V, Jiang X. 2022. Compositional and functional changes in microbial
1048 communities of composts due to the composting-related factors and the presence of *Listeria*
1049 *monocytogenes*. *Microbiol Spectr* **10**: e0184521. doi: 10.1128/spectrum.01845-21
1050 Wang Y, Luo L, Li Q, Wang H, Wang Y, Sun H, Xu J, Lan R, Ye C. 2019. Genomic dissection of
1051 the most prevalent *Listeria monocytogenes* clone, sequence type ST87, in China. *BMC*
1052 *Genomics* **20**: 1014. doi: 10.1186/s12864-019-6399-1
1053 Wishart DS, Han S, Saha S, Oler E, Peters H, Grant JR, Stothard P, Gautam V. 2023.
1054 PHASTEST: faster than PHASTER, better than PHAST. *Nucleic Acids Res* **51**: W443–
1055 W450. doi: 10.1093/nar/gkad382
1056 Wood DE, Salzberg SL. 2014. Kraken: ultrafast metagenomic sequence classification using
1057 exact alignments. *Genome Biol* **15**: R46. doi: 10.1186/gb-2014-15-3-r46
1058 Xie Z, Tang H. 2017. ISEScan: automated identification of insertion sequence elements in
1059 prokaryotic genomes. *Bioinformatics* **33**: 3340–3347. doi: 10.1093/bioinformatics/btx433
1060 Zhang N, Ma Z, Li D, Ni H, Sun B, Liang Y. 2022. Soil pH filters the association patterns of
1061 aluminum-tolerant microorganisms in rice paddies. *mSystems* **7**: e0102221. doi:
1062 10.1128/msystems.01022-21
1063 Zhao Q, Sun D, Tang X, Hou L, Liu M, Han P. 2023. An investigation of *Nitrospira* bacteria in
1064 coastal wetlands of China: distribution pattern and related environmental driving factors.
1065 *Front Mar Sci* **10**: 1288142. doi: 10.3389/fmars.2023.1288142
1066