Locally persistent lineages are an important component of zoonotic foodborne pathogen Shiga toxin-producing *Escherichia coli* O157:H7 disease ecology

Gillian A.M. Tarr,[1] William Finical,[1] Joshua M. Rounds,[2] Anna Panek,[3] Kirk Smith[2]

[1] Division of Environmental Health Sciences, School of Public Health, University of Minnesota, Minneapolis, MN, 55455

[2] Foodborne, Waterborne, Vectorborne, and Zoonotic Diseases Section, Minnesota Department of Health, St. Paul, MN, 55164

[3] Division of Epidemiology and Community Health, School of Public Health, University of Minnesota, Minneapolis, MN, 55455

Running title: STEC O157 locally persistent lineages

Corresponding author: Gillian A.M. Tarr, gtarr@umn.edu, 612-626-9308, MMC807, Room 1240 Mayo, 420 Delaware Street SE, Minneapolis, MN, 55455; ORCiDORCID: 0000-0001-7372-1034

Text (without Methods): 3500/3500

Abstract: 147/150

Abstract

The control of zoonotic foodborne pathogens, including Shiga toxin-producing *Escherichia coli* (STEC), has proven difficult, in part because we know remarkably little about how these pathogens are maintained in an ecosystem. In response to this gap, we have developed a method for identifying pathogen strains that persistently infect defined human populations, which we call locally persistent lineages (LPLs). Using STEC O157:H7 in Minnesota as a model, we show that LPLs are widespread across the O157 phylogenetic tree, contribute significantly to local disease burden, and can persist for several years. They are also spatially clustered in an area with large populations of cattle, considered O157 STEC's primary reservoir. Our results suggest that STEC O157:H7 LPLs are established in ecological niches that support persistence. Identifying LPLs, and ultimately their reservoirs, can advance understanding of their disease ecology and provide opportunities for disease control at both local and national levels.

Introduction

Zoonotic foodborne pathogens are particularly challenging to control.[1] For most such pathogens, including Shiga toxin-producing *Escherichia coli* (STEC), *Campylobacter jejuni*, and non-typhoidal *Salmonella*, their animal reservoirs are of prime importance in their disease ecology.[2] These pathogens can be transmitted fecal-orally through numerous routes, including direct animal contact; contact with contaminated environments; ingestion of contaminated water; person-to-person in limited chains; and, of course, via food.[3] Their prolonged survival in food and environmental media[4–8] means that infection may be transmitted via food transported across the country, or from contact with an animal in the case's backyard. This creates uncertainty in where an infection has come from, and thus how best to allocate resources to prevent infections.

Reported STEC incidence in the U.S. has increased >40% during the previous decade, more than almost any other zoonotic foodborne pathogen, and its estimated total disease burden has nearly tripled.[9–11] This is largely due to improvements in detection of non-O157 STEC,[11] though O157 STEC remains the greatest cause of hemolytic uremic syndrome, a severe form of acute kidney disease that disproportionately affects young children.[12,13] Simultaneously, new vehicles of infection, including produce and low-moisture foods,[14–16] have been recognized, further complicating attribution and disease control. Cattle and other ruminants are STEC's primary reservoirs, but the pathogen is so ubiquitous in these species that elimination from every ruminant population is not possible.[17,18]

What if, instead, we could target the reservoir populations responsible for the greatest disease burden? Unfortunately, our understanding of STEC disease ecology is not currently sufficient to identify those herds. Within a given human population, it is unknown what proportion of infections are attributable to imported food vs. local reservoirs, as most cases are never traced back to their source, and transmission from local reservoirs can ultimately occur through any route. Even among locally acquired infections, it is unknown what proportion of STEC strains persist long-term within a single reservoir population, as opposed to circulating within maintenance communities[19] or existing transitorily in an area before migrating to a new region via host movements. Understanding this disease ecology is central to identifying the reservoirs that pose the greatest risk and designing smartly targeted strategies to control them.

To begin delineating the details of STEC disease ecology, we previously introduced the concept of locally persistent lineages (LPLs), pathogen strains that successfully persist in an ecosystem, causing infections and giving rise to new variants.[20] LPLs provide a means to: 1) define strains persistently responsible for disease in the human population; and 2) identify their reservoirs. We found that infections with STEC O157:H7, the single most common STEC serotype recovered from people in North America, were associated with LPLs in 75% of cases reported in 2018 and 2019 in Alberta, Canada, a particularly high-incidence region with a substantial cattle population.[20] However, STEC disease ecology likely differs in areas with especially large reservoirs than in areas with fewer ruminants.

Here, we test the hypothesis that LPLs are a common feature of regional STEC disease dynamics and determine what role they play in local epidemiology, focusing on STEC O157:H7, the single most common and virulent serotype in the United States. We also generalize the LPL concept for use in areas where primarily only human case surveillance data are available; e.g., the entire

United States. In doing so, we identify regional variation in LPLs, contribute to the understanding of STEC disease ecology, and provide a framework for investigating reservoir dynamics for other major zoonotic foodborne pathogens.

Results

We identified 373 STEC O157:H7 isolates reported to the Minnesota Department of Health (MDH) from 2010 through 2019 that had been sequenced and were available on NCBI. Of these, 10 (2.6%) could not be downloaded, failed assembly, or failed QC. STECFinder confirmed the identity of the remaining 363 as STEC O157:H7. We randomly sampled 1,245 isolates external to Minnesota that had been reported and sequenced through PulseNet by other states, of which 52 (4.2%) could not be downloaded, failed assembly, or failed QC. After excluding a small divergent clade, 1,164 external isolates were included in analysis, for a final ratio of 3.2 external isolates to 1 Minnesota isolate.

Minnesota isolates were slightly overrepresented in 2010 and underrepresented during 2011-2016 (Supplemental Figure S1). Twelve percent of external isolates (n=141) were reported from neighboring Upper Midwest states (Iowa, South Dakota, North Dakota, Wisconsin). Clade F was the most common clade for both external and Minnesota isolates. The proportion of Minnesota isolates identified as clade E was three times that of external isolates (Supplemental Figure S2).

Across all 1,532 isolates, we identified 11,437 core genome SNPs. Prior to tree construction, we downsampled 116 MN (32.0%) and 354 external (30.4%) isolates due to high sequence similarity. The root of the time-calibrated phylogenetic tree was estimated to be 1840 (95% HPD 1800, 1875) (Figure 1). The estimated clock rate for the core genome was 8.33×10^{-5} (95% HPD 7.47×10^{-5}, 9.16×10^{-5}) substitutions/site/year. A reconstruction of the effective population size from the skyline analysis indicated a sharp increase beginning in the early 2000s (Supplemental Figure S3).

Locally Persistent Lineages Were Common

We identified 15 Minnesota STEC O157:H7 LPLs, defined as lineages associated with at least 3 human infections spread over 1 or more years, with more than ⅔ of all isolates in the lineage from Minnesota (Figure 1). These LPLs were associated with 35.3% (95% CI 30.3%, 40.4%) of Minnesota STEC O157:H7 cases reported during the study period, after reincorporating downsampled isolates (Table 1). We estimated incidence of 1.0 LPL-associated and 1.8 non-LPL-associated STEC O157:H7 cases per 100,000. LPLs were also associated with 21 external isolates, of which 10 were reported in other Upper Midwest states. Including both Minnesota and external cases, LPLs ranged in size from 3 to 23 isolates, with an average of 9.9 (SD 6.8). They produced cases for 1.3 to 8.6 years (mean 3.8, SD 2.8).

LPL Characteristics Differed by O157 Clade

LPLs occurred in all clades except A and B, neither of which included any Minnesota isolates (Figure 2). Clade G contained the greatest number of LPLs, with 7, followed by E and F, with 3 LPLs each. Clade G also had the greatest proportion of its Minnesota cases associated with LPLs (47.5%; 95% CI 38.3%, 56.8%) (Table 1). In Clade F, 29.8% (95% CI 22.6%, 37.8%) of Minnesota isolates were associated with LPLs, as were 25.6% (95% CI 16.4%, 36.8%) of Minnesota isolates in clade E. Clades C and D each contained 1 LPL with 3 Minnesota isolates each. The proportion of Minnesota isolates associated with LPLs differed significantly by clade ($p = 0.006$).

Clade F LPLs differed from those in other clades (Table 1). They had more isolates per LPL (mean 17.7, SD 5.9) and produced cases for more years (mean 8.1, SD 0.8); only the ANOVA comparing differences in mean duration across clades was statistically significant ($p = 0.008$). Clade F LPLs

also had a greater average maximum number of SNPs between isolates (46, SD 23) and earlier inferred ancestor (mean 2002, SD 4) compared to clades E and G.

LPL Isolates Differed from Non-LPL Isolates

A greater proportion of LPL than non-LPL isolates were linked to outbreaks; conversely, fewer LPL than non-LPL isolates were sporadic or not linked to known outbreaks (Table 2). The geographic source of the outbreaks (Minnesota or multi-state) differed by LPL status ($p < 0.001$). Forty (31.3%; 95% CI 23.4%, 40.0%) LPL isolates were linked to outbreaks limited to Minnesota or with Minnesota sources, compared to only 5.1% (95% CI 2.7%, 8.7%) of non-LPL isolates. Supporting the veracity of the LPLs, there were no LPL isolates linked to multi-state outbreaks, compared to 16.6% (95% CI 12.1%, 22.0%) of non-LPL isolates linked to multi-state outbreaks. Among isolates linked to Minnesota outbreaks, there was no difference between LPL and non-LPL isolates in the type of outbreak ($p = 0.596$).

LPL and non-LPL isolates also differed genetically (Table 2). There was a substantial difference in *stx* profile, a key virulence determinant ($p < 0.001$), with 39.1% (95% CI 30.6%, 48.1%) of LPL isolates carrying *stx1a/stx2a* combinations, compared to 13.2% (95% CI 9.1%, 18.2%) of non-LPL isolates. LPL isolates were less likely to carry only *stx2* genes, particularly when *stx2a* (6.3%; 95% CI 2.7%, 11.9%) or *stx2c* (1.6%; 95% CI 0.2%, 5.5%) occurred alone. These profiles were significantly more common (16.2% and 8.1%, respectively) among non-LPL isolates. Despite the differences in *stx* virulence profiles, there were no differences in clinical outcomes between cases infected with LPL vs. non-LPL STEC. LPL isolates were also less likely than non-LPL isolates to carry any antimicrobial resistance genes (ARGs) and the *qacE* gene, which is associated with tolerance to quaternary ammonium products. The frequency of other virulence genes was similar between LPL and non-LPL isolates.

We also tested isolates collected from animals or the environment. None of the 111 external or 4 Minnesota isolates clustered with LPLs (Supplemental Fig. S4).

Spatial Distribution of LPL and Non-LPL STEC Overlapped

We found geographic overlap in LPL and non-LPL cases (Figure 3A). Zip code tabulation areas (ZCTAs) with high rates of LPL cases were clustered in the western part of the state (Figure 3B). The incidence of LPL cases within this cluster was 2.97 times ($p = 0.001$) the incidence in all of Minnesota. We also found a non-LPL cluster in the southwest of the state, with a standardized incidence ratio (SIR) of 3.62 ($p = 0.001$). The two clusters overlapped, but both contained ZCTAs not part of the other cluster. However, the most likely cluster identified in a direct comparison of LPL to non-LPL cases was not statistically significant [relative risk (RR) 2.00, $p = 0.081$]. All results were consistent in sensitivity analyses of different window sizes.

Discussion

To better understand the disease ecology of zoonotic foodborne pathogens, we developed a method for distinguishing infections caused by persistent local strains, which we have termed locally persistent lineages (LPLs).[20] In this study, we identified 15 STEC O157:H7 LPLs in Minnesota, supporting our initial hypothesis that LPLs are a common feature of STEC disease systems. LPL-associated infections were epidemiologically important, accounting for 35% of all cases and 44% of cases linked to investigated outbreaks – more cases than were linked to multi-state outbreaks during the period.

Our findings support LPLs as genuine ecological phenomena linked to local reservoirs. While our LPL criteria required that more than two-thirds of isolates in a lineage had been reported in Minnesota, external isolates accounted for only 14% of LPL isolates, and almost half of those were reported in neighboring states in the Upper Midwest. This demonstrates even greater dominance by Minnesota isolates than expected. The LPLs contained no isolates linked to multi-state outbreaks, which if they had, would have suggested erroneous identification as LPLs. Similarly, none of the animal or environmental isolates from outside of Minnesota clustered with LPLs. While this provides evidence against the LPLs being erroneous, we had only 4 STEC O157:H7 isolates from animals and the environment in Minnesota, which were insufficient to link LPLs to local reservoirs. Both LPL and non-LPL spatial clusters were in areas of the state with high densities of domestic ruminants,[21] and the cluster of LPL-associated cases overlapped an area with a high concentration of dairy cattle (i.e., Minnesota's "dairy belt"). Further investigation is needed to determine whether LPLs are linked to dairy operations, but this clustering suggests an association between LPLs and a putative local reservoir. Additionally, the spatial clustering we found is broader than a single farm, indicating a role for the larger environment in STEC persistence. Overall, our findings are consistent with LPL-associated isolates originating from Minnesota reservoirs.

Across all STEC O157:H7 clades, 35% of Minnesota cases were associated with LPLs, which is lower than the 75% associated with LPLs in our previous study in Alberta.[20] This was not unexpected given the importance of cattle as a reservoir. Alberta has over twice the number of cattle, with 4.5 million head in 2020,[22] as Minnesota, with 2.1 million head in 2024.[57] It is well-established that living in a region with a high density of ruminants increases the risk of STEC.[21,23–28] LPLs could help explain the mechanism for this increased risk. Large ruminant populations are more likely than small populations to exceed the critical community size required for pathogen maintenance, resulting in less stochastic fade-out.[29–31] Strain persistence in ruminant-dense regions would provide greater opportunities for both direct and indirect human infections. However, a systematic study is needed to compare LPLs across geographic areas to identify regional factors that influence LPL-associated disease burden.

The proportion of LPL-associated cases in an area is also dependent on the sources of non-LPL infections. The absolute incidence of LPL vs. non-LPL O157 infections in Minnesota was 1.00 vs. 1.84 cases per 100,000 population, compared to 2.54 vs. 0.86 cases per 100,000 in Alberta, revealing a much greater burden of non-LPL-associated disease in Minnesota.[20] There are several possible origins of these other O157 infections, including infections acquired during travel, from imported food, or from non-persistent local sources. Approximately 4 to 9% of *E. coli* O157:H7 cases are believed to be associated with international travel,[9,10] leaving over half of Minnesota's

199 cases unattributed to either LPLs or travel. Infections caused by imported contaminated food likely
200 make up a large portion of this. Both Minnesota and Alberta have limited growing seasons and thus
201 import the majority of their produce, though the sources may differ.

202 The current study also suggests the existence of infections acquired locally but not associated with
203 LPLs. Non-LPL outbreaks traced to local sources demonstrate that these occur, and there is no
204 evidence that sporadic cases are not also attributable to such infections. Thirteen percent of
205 outbreak-linked cases were linked to Minnesota outbreaks but were not associated with LPLs. This
206 may be a slight overestimate, as none of the person-to-person outbreaks that MDH investigated
207 during the period were determined to be multi-state outbreaks, but the index case in these
208 outbreaks could have been infected by an external strain via imported food or travel. However,
209 there were still several outbreak cases not associated with LPLs that were linked to animal contact
210 in Minnesota or locally produced foods. The proportion of outbreak vs. sporadic cases linked to
211 LPLs were not identical, but they were similar, suggesting locally infected but non-LPL sporadic
212 cases exist just as outbreak cases do. The disease ecology of local, non-persistent O157 strains is
213 unclear. These could be strains passing transitorily through local reservoirs but unable to establish
214 lasting populations due to competition with strains better adapted to ecosystem-level persistence
215 or other factors. More work is needed to determine the role that such strains play in local disease
216 burden and the factors that enable some strains to persist while others do not.

217 We identified important genetic differences between LPL and non-LPL isolates. Of the three largest
218 clades in our current study, clade G had the greatest proportion of isolates associated with LPLs,
219 which is consistent with our findings in Alberta, where >90% of STEC O157:H7 isolates, including
220 all LPL isolates, belonged to clade G.[20] Future work is needed to determine whether there are
221 genetic factors within this clade that predispose its isolates to developing LPLs. The *stx1a/stx2a*
222 profile was also significantly more common among LPL isolates, and isolates with only an *stx2*
223 gene were more commonly not associated with LPLs. Virulence is attenuated in isolates with both
224 *stx1* and *stx2* genes relative to only *stx2*,[32–34] though we found no difference in clinical outcomes
225 between LPL- and non-LPL-associated cases. The *stx1/stx2* profile of LPL isolates may indicate
226 that high virulence is not an important trait for establishing or maintaining STEC populations in the
227 ecosystem. However, our work in Alberta found that LPLs adopted more virulent *stx* profiles over
228 time,[20] so it is also possible that selection on virulence is niche-specific. Our examination of other
229 virulence genes in the current study identified no differences between LPL and non-LPL isolates.

230 We found differences in genes for antimicrobial resistance and quaternary aluminum compound
231 (QAC) tolerance. In both cases, non-LPL isolates had significantly more resistance/tolerance
232 genes. Antimicrobial resistance has been associated with persistence within cattle,[35] and while
233 persistence within the animal host is a complex trait,[36,37] the reverse association between
234 resistance and LPLs suggests that the persistence of LPLs is not driven by intra-host persistence,
235 or if it is, other virulence factors are responsible. QAC tolerance in *E. coli* is a recognized issue and
236 is often coupled with antimicrobial resistance.[38] QACs are used regularly in the food industry for
237 washing produce and sanitizing surfaces used for processing animal products.[39,40] As a large
238 portion of non-LPL isolates likely originated from imported food products, whereas LPL infections
239 were likely spread by a mix of transmission routes, it is not unexpected that QAC tolerance is
240 present in non-LPL isolates but not in LPL isolates.

Our results suggesting persistence at the ecosystem level, rather than merely within individual hosts, is consistent with previous work. Challenge and natural infection studies show that STEC colonization and the bacterial load that animals shed decline rapidly after inoculation, persisting in most animals at low levels for <60 days.[41,42] Shedding longer than 2 months is seen in a small proportion of cattle, though this is often with serotypes not commonly associated with human infection.[43–45] Such short carriage times are inconsistent with LPL duration, which spanned up to 9 years in this study and 13 years in Alberta.[20] Longitudinal studies report large fluctuations in STEC prevalence,[46–48] and models of the role of the environment in maintaining STEC on farms predict fade-out in all cases.[49] Even accounting for the movement of cattle between farms, STEC persistence cannot be explained by cattle alone.[50,51] We have previously found that sheep density is also associated with STEC O157:H7 incidence,[21] suggesting that species other than cattle should also be modeled to determine their contributions to ecosystem-level persistence. The identification of LPLs is a starting point for investigating the other factors that enable STEC persistence within an ecosystem, as they provide targets that are well-established in an area for future study.

While identifying LPLs can advance understanding of STEC disease ecology, they also provide opportunities for disease control. Domestic ruminants are rarely tested for STEC, because they carry the pathogen largely without harm to themselves, and most serotypes are rarely, if ever, associated with human disease.[52,53] LPLs, however, are location-specific targets defined on the basis of repeated human infections. If their specific reservoir populations can be identified, control measures could be implemented to eradicate them or animals in those populations could be screened prior to slaughter. If an LPL strain is eradicated from a reservoir, the population should be monitored for new strains to avoid the establishment of new LPLs. Even without identifying the reservoir, the widespread adoption of whole genome sequencing means that public health authorities could identify cases associated with LPLs and direct their investigations to local sources. This could be particularly advantageous when resources for public health are limited. Despite using locality to define them, the impact of LPLs is national, as bacteria local to one area can contaminate food items that are shipped throughout the country.[54–57] Thus, controlling LPLs where they are local has the potential to prevent disease on a much larger scale.

To enable the identification of LPLs, the method we present here uses only standard surveillance data, whereas our previous LPL study relied upon region-wide sampling from the likely reservoir population over several years.[20] Such reservoir data is not widely available, as evidenced by the availability of only 4 animal or environmental isolates from Minnesota. Relying on host data also constrains the analysis to a specific hypothesized reservoir population. Conversely, the human surveillance data used in the present study remains agnostic to the composition of potential reservoirs and is already systematically collected by many countries for major zoonotic foodborne pathogens. However, whole genome sequencing has only emerged as a routine part of public health surveillance in the past decade, skewing the representativeness of sequenced isolates early in our study period. This precluded us from using phylogeographic methods,[58,59] which would have provided information about the source of non-LPL isolates.

Zoonotic foodborne pathogens are transmitted by several routes and innumerable vehicles. However, they all start at a reservoir, and each reservoir has an inherent locality. Despite intensive surveillance efforts in the human population, we know remarkably little about how zoonotic

foodborne pathogens with primary agricultural reservoirs are maintained in an ecosystem. Leveraging the local nature of reservoirs, we have developed a method for identifying strains with proven ability to infect humans year after year. The approach we present here uses only human surveillance data, which is more common than systematic surveys of potential reservoir species. Long persistence durations, genetic characteristics associated with lower virulence and resistance, and the spatial clustering of LPL cases suggest that STEC O157:H7 LPLs are established in ecological niches that support persistence, rather than merely being successful at persistently colonizing individual hosts. Identification of LPLs can advance understanding of their disease ecology and contribute to disease control, reducing disease burden both locally and nationally.

Online Methods

Study Population

We conducted a genomic epidemiology study of O157 STEC cases reported to: a) the Minnesota Department of Health (MDH) for Minnesota cases, or b) health departments outside of Minnesota for external cases. MDH and other health departments subsequently reported cases to the Centers for Disease Control and Prevention (CDC) PulseNet surveillance system.[60] Cases with an isolation date in 2010 through 2019 were included. Only cases reported within the 50 U.S. states with isolates that had been whole genome sequenced (WGS) were included. Cases were defined as individuals meeting the Council of State and Territorial Epidemiologists (CSTE) confirmed case definition that was in use at the time the case was reported; CSTE STEC case definitions applying to this study were adopted in 2006, 2014, and 2018.[61–64] In most instances, cases sought care and provided a fecal sample, which was tested by their clinician. Fecal samples, and in some cases bacterial isolates from culture, were delivered to public health laboratories, where, if necessary, fecal samples were cultured to obtain an isolate, and all isolates were whole genome sequenced per PulseNet procedures.[65] Public health laboratories participating in PulseNet adopted routine STEC sequencing between 2016 and 2019; sequences from isolates collected prior to routine sequencing were usually retrospectively sequenced due to particular interest in the case, such as involvement in an outbreak, severe outcome, or unique source of infection. MDH began routine STEC sequencing in 2016 and sequenced an additional 39 isolates from 2010 to 2015 to provide more data for the early years of the study.

MDH obtains extensive metadata about reported cases, including home zip code and whether a case was linked to a known outbreak. We developed our method to designate cases reported in Minnesota as “local” and cases reported in the other 49 states as “external”, though the code was designed to allow any state or region to be designated as the locality of interest. We included all available Minnesota O157 STEC sequences and randomly sampled external sequences at a ratio of approximately 3:1. Initial sampling was based on the serotype identified in the PulseNet database.

Sequence Assembly, Validation, and Characterization

We assembled all O157 STEC sequences from cases reported in Minnesota and the randomly sampled external sequences using the Bactopia v3.0.0 pipeline.[66] First, SRA Toolkit v3.0.0 was used to download all sequences from NCBI. Using Bactopia, FastQC v0.12.1 evaluated raw reads based on read count, sequence coverage, and sequence depth. Failed reads were excluded from subsequent assembly. We *de novo* assembled raw reads with Shovill v1.1.0 within Bactopia, which included use of Trimmomatic to trim adapters and read ends with quality <6. Reads with overall quality scores <10 after trimming were discarded. After Bactopia assembly, we assessed assembly quality based on number of contigs (≤400), genome size (≥5.0 Mb), N50 (>30,000), and L50 (≤50). Low-quality assemblies were excluded. We used STECFinder v1.1.025 to determine Shiga toxin gene (*stx*) profile and confirm serotype using the *wzx* or *wzy* O-antigen genes and detection of the H-antigen.[67] O157 STEC clades were assigned based on previously identified SNPs.[68] We detected other virulence genes, antimicrobial resistance genes (ARGs), and stress tolerance genes in Minnesota isolates with AMRFinder Plus v3.12.8 within the Bactopia pipeline,[69] using *Escherichia coli* O157:H7 str. Sakai (NC_002695.2)[70] as the reference genome and limiting results to matches

with >90% sequence coverage and identity. ARGs that did not correspond to a specific antibiotic, including efflux pump genes and *blaEC* were excluded from the analysis.

MDH identified 5 sequences of animal or environmental origin collected during 2010-2019 from their collection. We identified 115 STEC O157:H7 animal or environmental isolates that had been reported to PulseNet during the study period from the other 49 states. All sequences were available on NCBI. Four and 111 sequences were successfully assembled and passed QC.

Sequence Alignment

To avoid inferring deeper evolutionary histories than our data could support, we excluded isolates in clade A (n=29), which included only external isolates. No clade B isolates from Minnesota passed QC; however, because clade B was not excluded *a priori* due to divergence, external clade B isolates were left in the analysis. We generated a core SNP alignment with recombinant blocks removed using Bactopia's Snippy workflow, which incorporates Snippy v4.6.0, Gubbins v3.3.0, and IQTree v2.2.2.7, followed by SNP-Sites v2.5.1.[66,71] The number of core SNPs between isolates was calculated using PairSNP v0.3.1. Prior to phylogenetic analysis, we pruned alignments of similar sequences agnostic to location in a process called "downsampling".[20] Downsampled sequences were within 2 SNPs of another sequence, referred to as the matching sequence. Additionally, downsampled sequences were from cases reported <60 days after a matching sequence, with no more than 12 months of continuous cases from the initial matching sequence (Supplemental Fig. S5). We noted the matching sequence retained within the alignment to later incorporate the downsampled sequences into the analysis of LPLs. If any isolate within a downsampled group was linked to a known outbreak, all isolates in the downsampled group were considered to have been part of the same outbreak.

Animal and environmental isolates were examined by rerunning Snippy and Gubbins with all human, animal, and environmental isolates. We created an IQTree to examine the relationships between isolates and visualized the tree in Microreact v282 with metadata for LPL, source, and geographic location.

Phylogenetic Analysis

We used BEAST2 v2.6.7 to generate a time-calibrated phylogeny, which is a phylogenetic tree on the scale of time.[72] To allow for changes in population dynamics over time, we used a Bayesian coalescent skyline tree model with 3 skyline pieces. For the substitution model, we used an HKY prior with empirical frequencies and discrete Gamma site model with four categories. For the clock model we used a relaxed log-normal prior with initial value 1.5×10^{-5} and log-normal distribution ($M=1.5\times10^{-5}$, S=1.5, with mean in real space).[20] Four chains were run till convergence with >200 estimated sample size for all non-skyline parameters. A maximum clade credibility (MCC) tree was generated from the combined chains.

LPLs were identified based on the following criteria: 1) a single lineage of the MCC tree with a most recent common ancestor (MRCA) with ≥80% posterior probability; 2) contained ≥3 sequences; 3) contained >⅔ Minnesota sequences; 4) sequences spanned a period of ≥1 year; and 5) all isolates were ≤100 core SNPs from one another. LPLs were highlighted on phylogenies for each genomic cluster using the treeio and ggtree packages in R.[73–75]

Statistical Analysis

LPL-associated cases and incidence

We calculated the proportion of Minnesota and external isolates, separately, that were associated with LPLs, along with Clopper-Pearson exact binomial 95% CIs. We calculated the incidence rate of LPL- and non-LPL-associated O157 STEC cases in Minnesota from 2010-2019, by multiplying the proportion associated with an LPL, or 1 minus that proportion for the non-LPL rate, by the overall STEC incidence rate (7.08 per 100,000[13]), multiplied by the proportion of reported STEC cases in the state that were infected with STEC O157:H7 (40.2%[21]). We compared these incidence rates to similarly calculated rates for Alberta, Canada, where LPLs have previously been investigated: LPL proportion (74.7%[20]) times overall STEC incidence (8.45 per 100,000[76]) times the proportion of cases that were infected with STEC O157:H7 (40.3%[77]).

Differences by clade and LPL status

We compared the proportion of cases associated with LPLs across clades using a chi-square test. For each LPL, we determined the number of isolates per LPL, the maximum number of SNPs between sequences, year of its MRCA, and duration from earliest isolate to latest isolate. For these measures, we calculated the mean value and standard deviation (SD) for all O157 STEC and for each clade. We compared the means between clades using analysis of variance (ANOVA), limiting the analysis to clades with >1 LPL. If the Shapiro-Wilk test indicated non-normality in ANOVA residuals, we investigated the source (e.g. an outlier) and performed a sensitivity analysis (e.g. without the outlier).

We calculated the frequency of LPL vs. non-LPL isolates that were associated with outbreaks investigated by MDH; had bloody diarrhea, were hospitalized, or had HUS; and carried each combination of *stx* genes, other virulence genes, ARGs, and stress tolerance genes. For each proportion, we calculated an exact binomial 95% CI and used a chi-square test for independence or Fisher's exact test to compare proportions between LPL and non-LPL groups. All analyses were performed in R.[73]

LPL spatial distribution

To determine whether LPL isolates were geographically clustered, as would be expected if they represent local patterns of disease ecology, we assessed the spatial distribution of LPL and non-LPL isolates. We assigned each isolate to the ZCTA of the reported case's residence and created a bivariate choropleth map using the biscale package.[78] To statistically test for spatial clustering, we used Poisson spatial scan statistics in the SpatialEpi package[79] to identify separate spatial clusters of LPL and non-LPL isolates. We used the population within each ZCTA as the reference population and limited clusters to containing up to 10% of the population. Sensitivity analyses adjusted this limit to 20% and 50% of the population. The SIR of each spatial cluster gave the ratio of the LPL or non-LPL incidence rate within the cluster relative to the state average. We also directly compared LPL to non-LPL isolates using Bernoulli scan statistics in the smacpod package,[80] which identified spatial clusters in which the ratio of LPL to non-LPL isolates was statistically greater inside the cluster than outside the cluster, given as a RR. The maximum cluster size was set at a radius of 50 miles; 25 and 100 mile radii were tested during sensitivity analysis. The alpha level for the primary cluster in all analyses was set at 0.05.

Acknowledgements

We thank Phillip I. Tarr, Craig Hedberg, Jeffrey Bender, and Noelle Noyes for their contributions to the study and feedback on the manuscript. We gratefully acknowledge the PulseNet participating laboratories, whose data was used for the creation of this publication. This research was funded in part by NIH/NIAID K01AI168499. The funder had no role in study design, data collection and interpretation, or the decision to submit the work for publication. The authors acknowledge the Minnesota Supercomputing Institute (MSI) at the University of Minnesota for providing resources that contributed to the research reported in this paper.

428 ## References

429 1. Hassan, A. *et al.* Emerging Foodborne Pathogens: Challenges and Strategies for Ensuring Food
430 Safety. *Biol. Life Sci. Forum* **31**, 32 (2023).
431 2. Fenton, A. & Pedersen, A. B. Community epidemiology framework for classifying disease
432 threats. *Emerg. Infect. Dis.* **11**, 1815–1821 (2005).
433 3. DuPont, H. L. The Growing Threat of Foodborne Bacterial Enteropathogens of Animal Origin.
434 *Clin. Infect. Dis.* **45**, 1353–1361 (2007).
435 4. Durso, L. M., Gilley, J. E. & Miller, D. N. Differential Survival of Non-O157 Shiga Toxigenic
436 Escherichia coli in Simulated Cattle Feedlot Runoff. *Foodborne Pathog. Dis.* **18**, 771–777
437 (2021).
438 5. Kim, J. & Jiang, X. The growth potential of Escherichia coli O157:H7, Salmonella spp. and
439 Listeria monocytogenes in dairy manure-based compost in a greenhouse setting under
440 different seasons. *J. Appl. Microbiol.* **109**, 2095–2104 (2010).
441 6. Bardsley, C. A. *et al.* Strain, Soil-Type, Irrigation Regimen, and Poultry Litter Influence
442 Salmonella Survival and Die-off in Agricultural Soils. *Front. Microbiol.* **12**, (2021).
443 7. Alegbeleye, O. & Sant'Ana, A. S. Survival of *Salmonella* spp. under varying temperature and soil
444 conditions. *Sci. Total Environ.* **884**, 163744 (2023).
445 8. Good, L. *et al.* Strain-Specific Differences in Survival of Campylobacter spp. in Naturally
446 Contaminated Turkey Feces and Water. *Appl. Environ. Microbiol.* **85**, e01579-19 (2019).
447 9. Scallan, E. *et al.* Foodborne Illness Acquired in the United States-Major Pathogens. *Emerg.*
448 *Infect. Dis.* **17**, 7–15 (2011).
449 10. Scallan Walter, E. J. *et al.* Foodborne Illness Acquired in the United States—Major Pathogens,
450 2019 - Volume 31, Number 4—April 2025 - Emerging Infectious Diseases journal - CDC.
451 doi:10.3201/eid3104.240913.

452 11. Shah, H. J. Reported Incidence of Infections Caused by Pathogens Transmitted Commonly
453 Through Food: Impact of Increased Use of Culture-Independent Diagnostic Tests — Foodborne
454 Diseases Active Surveillance Network, 1996–2023. *MMWR Morb. Mortal. Wkly. Rep.* **73**, (2024).
455 12. FoodNet Fast | CDC.
456 https://web.archive.org/web/20250201211029/https://wwwn.cdc.gov/FoodNetFast/HUS
457 (2025).
458 13. Centers for Disease Control and Prevention. FoodNet Fast. *Foodborne Diseases Active*
459 *Surveillance Network* (2023).
460 14. Carstens, C. K., Salazar, J. K. & Darkoh, C. Multistate Outbreaks of Foodborne Illness in the
461 United States Associated With Fresh Produce From 2010 to 2017. *Front. Microbiol.* **10**, 2667
462 (2019).
463 15. Crowe, S. J. *et al.* Shiga Toxin-Producing E. coli Infections Associated with Flour. *N Engl J Med*
464 **377**, 2036–2043 (2017).
465 16. Miller, B. D. *et al.* Use of traceback methods to confirm the source of a multistate Escherichia
466 coli O157:H7 outbreak due to in-shell hazelnuts. *J Food Prot* **75**, 320–7 (2012).
467 17. Mughini-Gras, L. *et al.* Attribution of human infections with Shiga toxin-producing Escherichia
468 coli (STEC) to livestock sources and identification of source-specific risk factors, The
469 Netherlands (2010-2014). *Zoonoses Public Health* **65**, e8–e22 (2018).
470 18. Smith, J. L., Fratamico, P. M. & Gunther, N. W. 4th. Shiga toxin-producing Escherichia coli. in
471 *Advances in applied microbiology* vol. 86 145–197 (United States, 2014).
472 19. Haydon, D. T., Cleaveland, S., Taylor, L. H. & Laurenson, M. K. Identifying reservoirs of
473 infection: a conceptual and practical challenge. *Emerg. Infect. Dis.* **8**, 1468–1473 (2002).

20. Tarr, G. A. *et al.* Persistent cross-species transmission systems dominate Shiga toxin-producing Escherichia coli O157:H7 epidemiology in a high incidence region: A genomic epidemiology study. *eLife* **13**, RP97643 (2025).
21. Ward, C. *et al.* Ruminant-dense environments increase risk of reported Shiga toxin-producing Escherichia coli infections independently of ruminant contact. *Appl. Environ. Microbiol.* e01864-24 (2025) doi:10.1128/aem.01864-24.
22. Livestock statistics | Alberta.ca. https://www.alberta.ca/livestock-statistics (2025).
23. Ohaiseadha, C., Hynds, P. D., Fallon, U. B. & O'Dwyer, J. A geostatistical investigation of agricultural and infrastructural risk factors associated with primary verotoxigenic E. coli (VTEC) infection in the Republic of Ireland, 2008-2013. *Epidemiol Infect* **145**, 95–105 (2017).
24. Jaros, P. *et al.* A prospective case-control and molecular epidemiological study of human cases of Shiga toxin-producing Escherichia coli in New Zealand. *BMC Infect Dis* **13**, 450 (2013).
25. Elson, R. *et al.* A spatial and temporal analysis of risk factors associated with sporadic Shiga toxin-producing Escherichia coli O157 infection in England between 2009 and 2015. *Epidemiol. Infect.* **146**, 1928–1939 (2018).
26. Frank, C., Kapfhammer, S., Werber, D., Stark, K. & Held, L. Cattle density and Shiga toxin-producing Escherichia coli infection in Germany: increased risk for most but not all serogroups. *Vector Borne Zoonotic Dis. Larchmt. N* **8**, 635–643 (2008).
27. Friesema, I. H., Van De Kassteele, J., D. E. Jager CM, Heuvelink, A. E. & Van Pelt, W. Geographical association between livestock density and human Shiga toxin-producing Escherichia coli O157 infections. *Epidemiol Infect* **139**, 1081–7 (2011).
28. Mulder, A. C. *et al.* Spatial Effects of Livestock Farming on Human Infections With Shiga Toxin-Producing Escherichia coli O157 in Small but Densely Populated Regions: The Case of the Netherlands. *GeoHealth* **4**, e2020GH000276 (2020).

498 29. Wood, J. C., McKendrick, I. J. & Gettinby, G. A simulation model to assess herd-level
499 intervention strategies against E. coli O157. *Epidemiol Infect* **135**, 749–64 (2007).
500 30. Zhang, X. S., Chase-Topping, M. E., McKendrick, I. J., Savill, N. J. & Woolhouse, M. E. Spread of
501 E. coli O157 infection among Scottish cattle farms: stochastic models and model selection.
502 *Epidemics* **2**, 11–20 (2010).
503 31. Zhang, X. S. & Woolhouse, M. E. Escherichia coli O157 infection on Scottish cattle farms:
504 dynamics and control. *J R Soc Interface* **8**, 1051–8 (2011).
505 32. Tarr, G. A. M. *et al.* Contribution and Interaction of Shiga Toxin Genes to Escherichia coli
506 O157:H7 Virulence. *Toxins* **11**, (2019).
507 33. Orth, D. *et al.* The Shiga toxin genotype rather than the amount of Shiga toxin or the cytotoxicity
508 of Shiga toxin in vitro correlates with the appearance of the hemolytic uremic syndrome. *Diagn*
509 *Microbiol Infect Dis* **59**, 235–42 (2007).
510 34. Petro, C. D. *et al.* Shiga Toxin Type 1a (Stx1a) Reduces the Toxicity of the More Potent Stx2a In
511 Vivo and In Vitro. *Infect Immun* **87**, (2019).
512 35. Salaheen, S. *et al.* Genomic diversity of antimicrobial-resistant and Shiga toxin gene-harboring
513 non-O157 *Escherichia coli* from dairy calves. *J. Glob. Antimicrob. Resist.* **33**, 164–170 (2023).
514 36. Barth, S. A. *et al.* The Accessory Genome of Shiga Toxin-Producing Escherichia coli Defines a
515 Persistent Colonization Type in Cattle. *Appl. Environ. Microbiol.* **82**, 5455–5464 (2016).
516 37. Barth, S. A. *et al.* Metabolic Traits of Bovine Shiga Toxin-Producing Escherichia Coli (STEC)
517 Strains with Different Colonization Properties. *Toxins Basel* **12**, (2020).
518 38. Soumet, C. *et al.* Reduced susceptibilities to biocides and resistance to antibiotics in
519 food-associated bacteria following exposure to quaternary ammonium compounds. *J. Appl.*
520 *Microbiol.* **121**, 1275–1281 (2016).

521 39. Chaidez, C., Lopez, J. & Castro-del Campo, N. Quaternary ammonium compounds: an
522 alternative disinfection method for fresh produce wash water. *J. Water Health* **5,** 329–333
523 (2007).
524 40. Jiang, X. *et al.* Characterization and horizontal transfer of qacH-associated class 1 integrons in
525 Escherichia coli isolated from retail meats. *Int. J. Food Microbiol.* **258,** 12–17 (2017).
526 41. Sanderson, M. W., Besser, T. E., Gay, J. M., Gay, C. C. & Hancock, D. D. Fecal Escherichia coli
527 O157:H7 shedding patterns of orally inoculated calves. *Vet. Microbiol.* **69,** 199–205 (1999).
528 42. Widiasih, D. A., Ido, N., Omoe, K., Sugii, S. & Shinagawa, K. Duration and magnitude of faecal
529 shedding of Shiga toxin-producing Escherichia coli from naturally infected cattle. *Epidemiol.*
530 *Infect.* **132,** 67–75 (2004).
531 43. Lim, J. Y. *et al.* Escherichia coli O157:H7 Colonization at the Rectoanal Junction of Long-
532 Duration Culture-Positive Cattle. *Appl. Environ. Microbiol.* **73,** 1380–1382 (2007).
533 44. Menrath, A. *et al.* Shiga toxin producing Escherichia coli: identification of non-O157:H7-Super-
534 Shedding cows and related risk factors. *Gut Pathog.* **2,** 7 (2010).
535 45. Wang, L. Y. R. *et al.* Multi-Year Persistence of Verotoxigenic Escherichia coli (VTEC) in a Closed
536 Canadian Beef Herd: A Cohort Study. *Front. Microbiol.* **9,** 2040 (2018).
537 46. Geue, L. *et al.* A long-term study on the prevalence of shiga toxin-producing Escherichia coli
538 (STEC) on four German cattle farms. *Epidemiol Infect* **129,** 173–85 (2002).
539 47. LeJeune, J. T. *et al.* Longitudinal study of fecal shedding of Escherichia coli O157:H7 in feedlot
540 cattle: predominance and persistence of specific clonal types despite massive cattle
541 population turnover. *Appl Env. Microbiol* **70,** 377–84 (2004).
542 48. Matthews, L. *et al.* Super-shedding cattle and the transmission dynamics of Escherichia coli
543 O157. *Epidemiol Infect* **134,** 131–42 (2006).

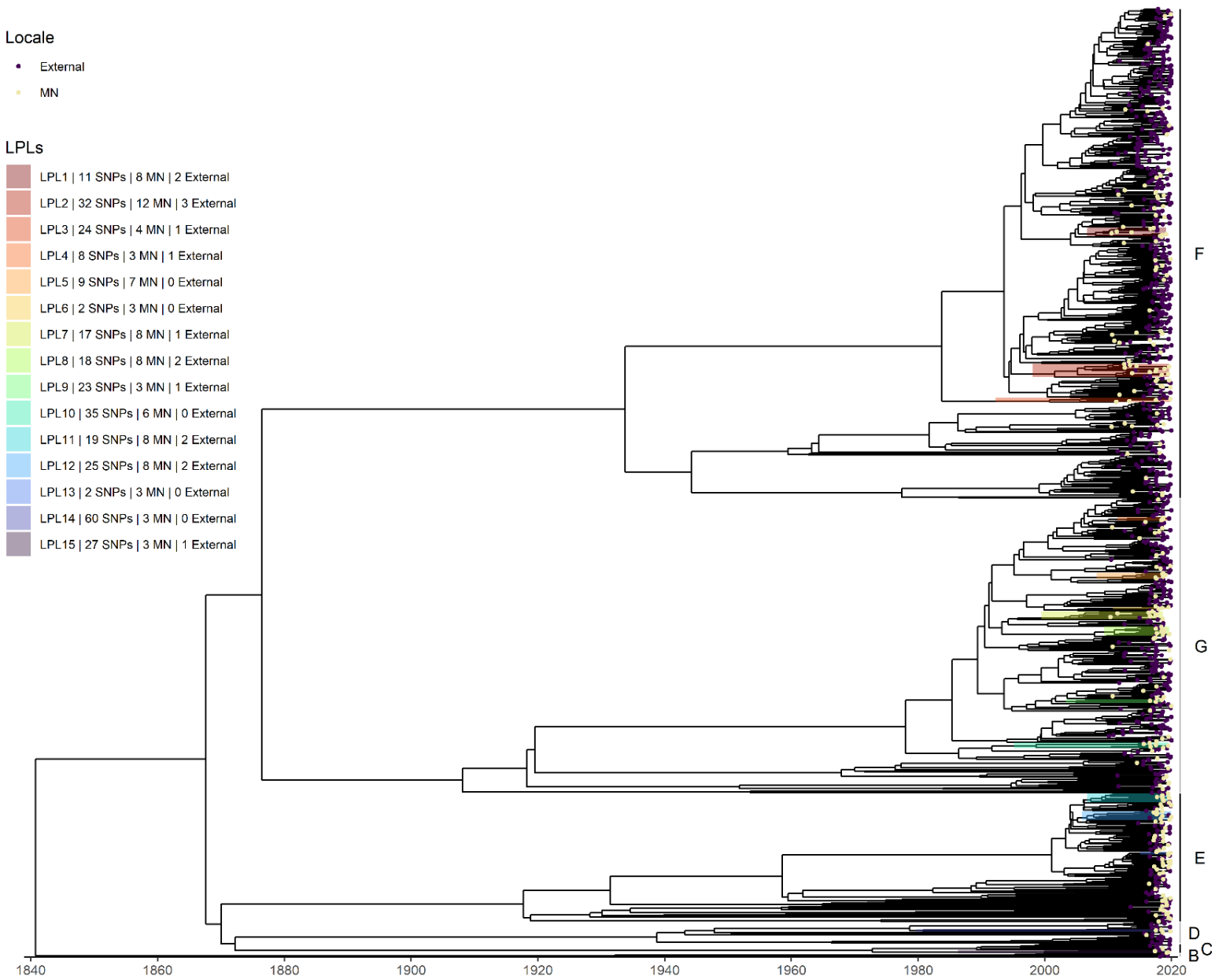
544 49. Wang, X., Gautam, R., Pinedo, P. J., Allen, L. J. & Ivanek, R. A stochastic model for
545 transmission, extinction and outbreak of Escherichia coli O157:H7 in cattle as affected by
546 ambient temperature and cleaning practices. *J Math Biol* **69**, 501–32 (2014).
547 50. Liu, W. C. *et al.* Metapopulation dynamics of Escherichia coli O157 in cattle: an exploratory
548 model. *J R Soc Interface* **4**, 917–24 (2007).
549 51. Stacey, K. F., Parsons, D. J., Christiansen, K. H. & Burton, C. H. Assessing the effect of
550 interventions on the risk of cattle and sheep carrying Escherichia coli O157:H7 to the abattoir
551 using a stochastic model. *Prev Vet Med* **79**, 32–45 (2007).
552 52. Persad, A. K. & LeJeune, J. T. Animal Reservoirs of Shiga Toxin-Producing Escherichia coli.
553 *Microbiol. Spectr.* **2**, 10.1128/microbiolspec.ehec-0027–2014 (2014).
554 53. Patterson, L., Navarro-Gonzalez, N., Jay-Russell, M. T., Aminabadi, P. & Pires, A. F. A. Risk
555 factors of Shiga toxin-producing *Escherichia coli* in livestock raised on diversified small-scale
556 farms in California. *Epidemiol. Infect.* **150**, e125 (2022).
557 54. Human Foods Program. Delmarva Produce Safety Environmental Research Study (2009 –
558 2014). *FDA* (2024).
559 55. Centers for Disease, C. & Prevention. Outbreak of Escherichia coli O157:H7 infections
560 associated with drinking unpasteurized commercial apple juice--British Columbia, California,
561 Colorado, and Washington, October 1996. *MMWR Morb Mortal Wkly Rep* **45**, 975 (1996).
562 56. Viazis, S. *et al.* An Overview of Farm Investigation Findings Associated with Outbreaks of Shiga
563 Toxin-Producing *Escherichia coli* Infections Linked to Leafy Greens: 2009 - 2021. *J. Food Prot.*
564 100542 (2025) doi:10.1016/j.jfp.2025.100542.
565 57. Centers for Disease Control & Prevention. Ongoing multistate outbreak of *Escherichia coli*
566 serotype O157:H7 infections associated with consumption of fresh spinach--United States,
567 September 2006. *MMWR Morb Mortal Wkly Rep* **55**, 1045–6 (2006).

568 58. Muller, N. F., Rasmussen, D. & Stadler, T. MASCOT: parameter and state inference under the
569 marginal structured coalescent approximation. *Bioinformatics* **34**, 3843–3848 (2018).
570 59. Franz, E. *et al.* Phylogeographic Analysis Reveals Multiple International transmission Events
571 Have Driven the Global Emergence of Escherichia coli O157:H7. *Clin Infect Dis* **69**, 428–437
572 (2019).
573 60. CDC. About PulseNet. *PulseNet* https://www.cdc.gov/pulsenet/hcp/about/index.html (2024).
574 61. Council of State and Territorial Epidemiologists. Public Health Reporting and National
575 Notification for Shiga Toxin-Producing *Escherichia coli* (STEC). (2017).
576 62. Council of State and Territorial Epidemiologists. Update to Public Health Reporting for Shiga
577 toxin-producing Escherichia coli (STEC). (2013).
578 63. Council of State and Territorial Epidemiologists. Public Health Reporting and National
579 Notification for Escherichia coli, Shiga toxin-producing (STEC). (2009).
580 64. Council of State and Territorial Epidemiologists. Revision of the Enterohemorrhagic
581 Escherichia coli (EHEC) condition name to Shiga toxin-producing Escherichia coli (STEC) and
582 Adoption of serotype specific national reporting for STEC. (2005).
583 65. Ribot, E. M., Freeman, M., Hise, K. B. & Gerner-Smidt, P. PulseNet: Entering the Age of Next-
584 Generation Sequencing. *Foodborne Pathog. Dis.* **16**, 451–456 (2019).
585 66. Petit, R. A. & Read, T. D. Bactopia: a Flexible Pipeline for Complete Analysis of Bacterial
586 Genomes. *mSystems* **5**, (2020).
587 67. Zhang, X., Payne, M., Kaur, S. & Lan, R. Improved Genomic Identification, Clustering, and
588 Serotyping of Shiga Toxin-Producing Escherichia coli Using Cluster/Serotype-Specific Gene
589 Markers. *Front Cell Infect Microbiol* **11**, 772574 (2021).
590 68. Strachan, N. J. *et al.* Whole Genome Sequencing demonstrates that Geographic Variation of
591 Escherichia coli O157 Genotypes Dominates Host Association. *Sci Rep* **5**, 14145 (2015).

592 69. Feldgarden, M. *et al.* AMRFinderPlus and the Reference Gene Catalog facilitate examination of
593 the genomic links among antimicrobial resistance, stress response, and virulence. *Sci. Rep.*
594 **11**, 12728 (2021).
595 70. Escherichia coli O157:H7 str. Sakai genome assembly ASM886v2. *NCBI*
596 https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000008865.2/.
597 71. Seemann, T. Snippy: fast bacterial variant calling from NGS reads. (2020).
598 72. Bouckaert, R. *et al.* BEAST 2.5: An advanced software platform for Bayesian evolutionary
599 analysis. *PLoS Comput Biol* **15**, e1006650 (2019).
600 73. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for
601 Statistical Computing (2022).
602 74. Wang, L.-G. *et al.* Treeio: An R Package for Phylogenetic Tree Input and Output with Richly
603 Annotated and Associated Data. *Mol. Biol. Evol.* **37**, 599–603 (2020).
604 75. Yu, G., Smith, D. K., Zhu, H., Guan, Y. & Lam, T. T.-Y. ggtree: an r package for visualization and
605 annotation of phylogenetic trees with their covariates and other associated data. *Methods*
606 *Ecol. Evol.* **8**, 28–36 (2017).
607 76. Interactive Health Data Application - Display Results.
608 http://www.ahw.gov.ab.ca/IHDA_Retrieval/selectSubCategoryParameters.do.
609 77. Glassman, H., Ferrato, C. & Chui, L. Epidemiology of Non-O157 Shiga Toxin-Producing
610 Escherichia coli in the Province of Alberta, Canada, from 2018 to 2021. *Microorganisms* **10**, 814
611 (2022).
612 78. Prener, C., Grossenbacher, T. & Zehr, A. biscale: Tools and Palettes for Bivariate Thematic
613 Mapping. (2022).
614 79. Kim, A. Y., Wakefield, J. & Moise, M. SpatialEpi: Methods and Data for Spatial Epidemiology.
615 1.2.8 https://doi.org/10.32614/CRAN.package.SpatialEpi (2012).

80. French, J. smacpod: Statistical Methods for the Analysis of Case-Control Point Data. 2.6.4 https://doi.org/10.32614/CRAN.package.smacpod (2014).

619 # Figures

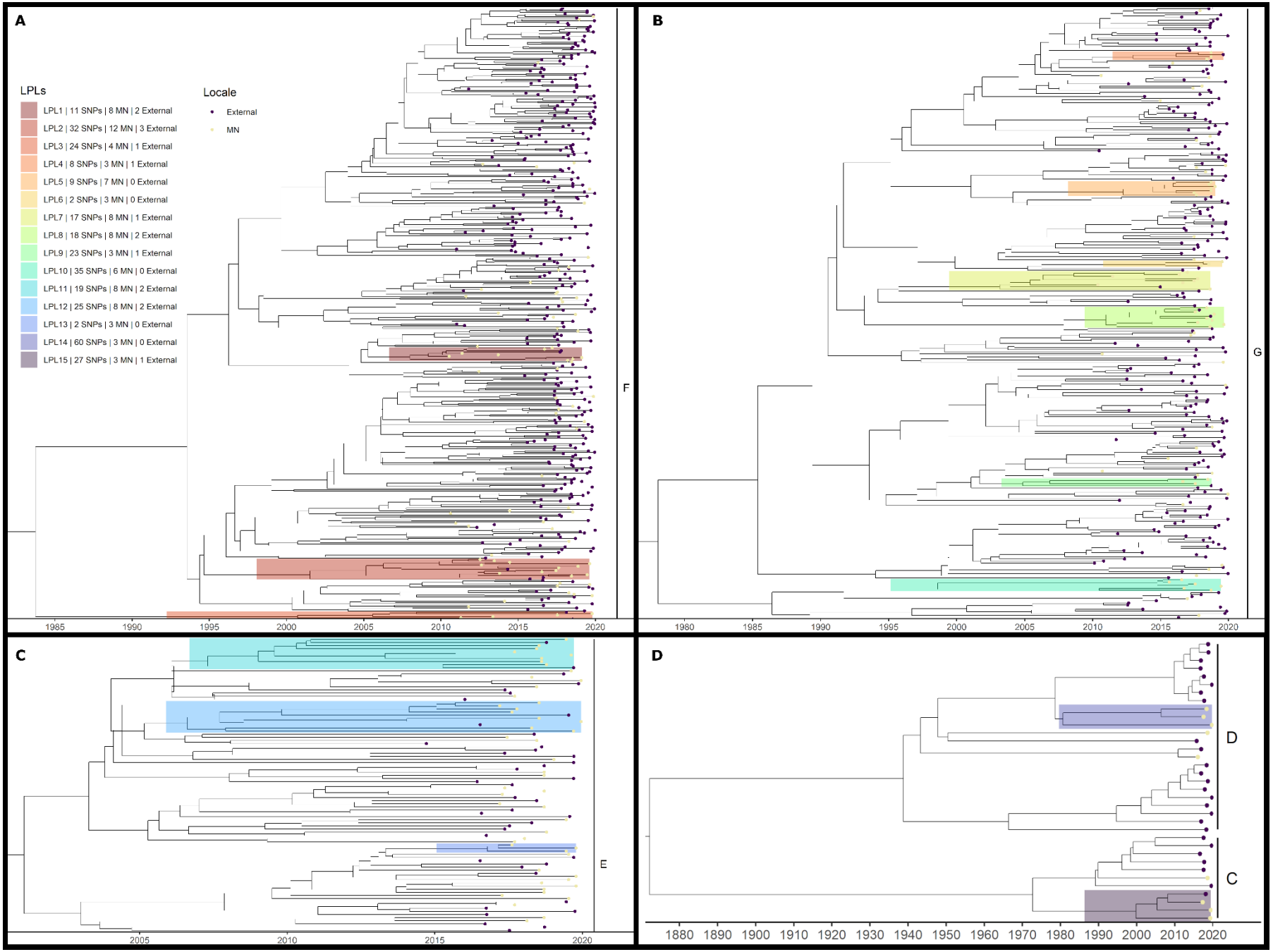

Locale
External
MN
LPLs
LPL1 | 11 SNPs | 8 MN | 2 External
LPL2 | 32 SNPs | 12 MN | 3 External
LPL3 | 24 SNPs | 4 MN | 1 External
LPL4 | 8 SNPs | 3 MN | 1 External
LPL5 | 9 SNPs | 7 MN | 0 External
LPL6 | 2 SNPs | 3 MN | 0 External
LPL7 | 17 SNPs | 8 MN | 1 External
LPL8 | 18 SNPs | 8 MN | 2 External
LPL9 | 23 SNPs | 3 MN | 1 External
LPL10 | 35 SNPs | 6 MN | 0 External
LPL11 | 19 SNPs | 8 MN | 2 External
LPL12 | 25 SNPs | 8 MN | 2 External
LPL13 | 2 SNPs | 3 MN | 0 External
LPL14 | 60 SNPs | 3 MN | 0 External
LPL15 | 27 SNPs | 3 MN | 1 External
F
G
E
D
C
B
1840
1860
1880
1900
1920
1940
1960
1980
2000
2020

620

621 **Figure 1. The STEC O157:H7 phylogeny and its LPLs.** We identified 15 LPLs, which included 128 MN (35.3%) and 21 external (1.8%)
622 isolates. The number of isolates used to define the LPLs is shown in the legend; the inclusion of downsampled isolates is not shown on
623 the figure but is included in the total. LPLs occurred in all clades except clade B.

624

625


A
LPLs
LPL1 | 11 SNPs | 8 MN | 2 External
LPL2 | 32 SNPs | 12 MN | 3 External
LPL3 | 24 SNPs | 4 MN | 1 External
LPL4 | 8 SNPs | 3 MN | 1 External
LPL5 | 9 SNPs | 7 MN | 0 External
LPL6 | 2 SNPs | 3 MN | 0 External
LPL7 | 17 SNPs | 8 MN | 1 External
LPL8 | 18 SNPs | 8 MN | 2 External
LPL9 | 23 SNPs | 3 MN | 1 External
LPL10 | 35 SNPs | 6 MN | 0 External
LPL11 | 19 SNPs | 8 MN | 2 External
LPL12 | 25 SNPs | 8 MN | 2 External
LPL13 | 2 SNPs | 3 MN | 0 External
LPL14 | 60 SNPs | 3 MN | 0 External
LPL15 | 27 SNPs | 3 MN | 1 External
Locale
External
MN
F
1985
1990
1995
2000
2005
2010
2015
2020
B
G
1980
1985
1990
1995
2000
2005
2010
2015
2020
C
E
2005
2010
2015
2020
D
D
C
1880
1890
1900
1910
1920
1930
1940
1950
1960
1970
1980
1990
2000
2010
2020

626

627 **Figure 2. Individual clades and their LPLs.** The number of isolates used to define the LPLs is shown in the legend; the inclusion of
628 downsampled isolates is not included in the figure but is included in the totals. A) Clade F contained 3 LPLs, accounting for 45 of its 151
629 Minnesota isolates (29.8%). LPLs in this clade produced cases over an average of 8.1 years. B) Clade G contained 7 LPLs, accounting for
630 57 of its 120 Minnesota isolates (47.5%). LPLs in this clade produced cases over an average of 3.0 years. C) Clade E contained 3 LPLs,
631 accounting for 20 of its 78 Minnesota isolates (25.6%). LPLs in this clade produced cases over an average of 2.5 years. D) Clades C and D
632 each contained 1 LPL. In clade C, this accounted for 3 of 8 Minnesota isolates (37.5%), and in clade D, it accounted for 3 of 6 Minnesota
633 isolates (50.0%).

634

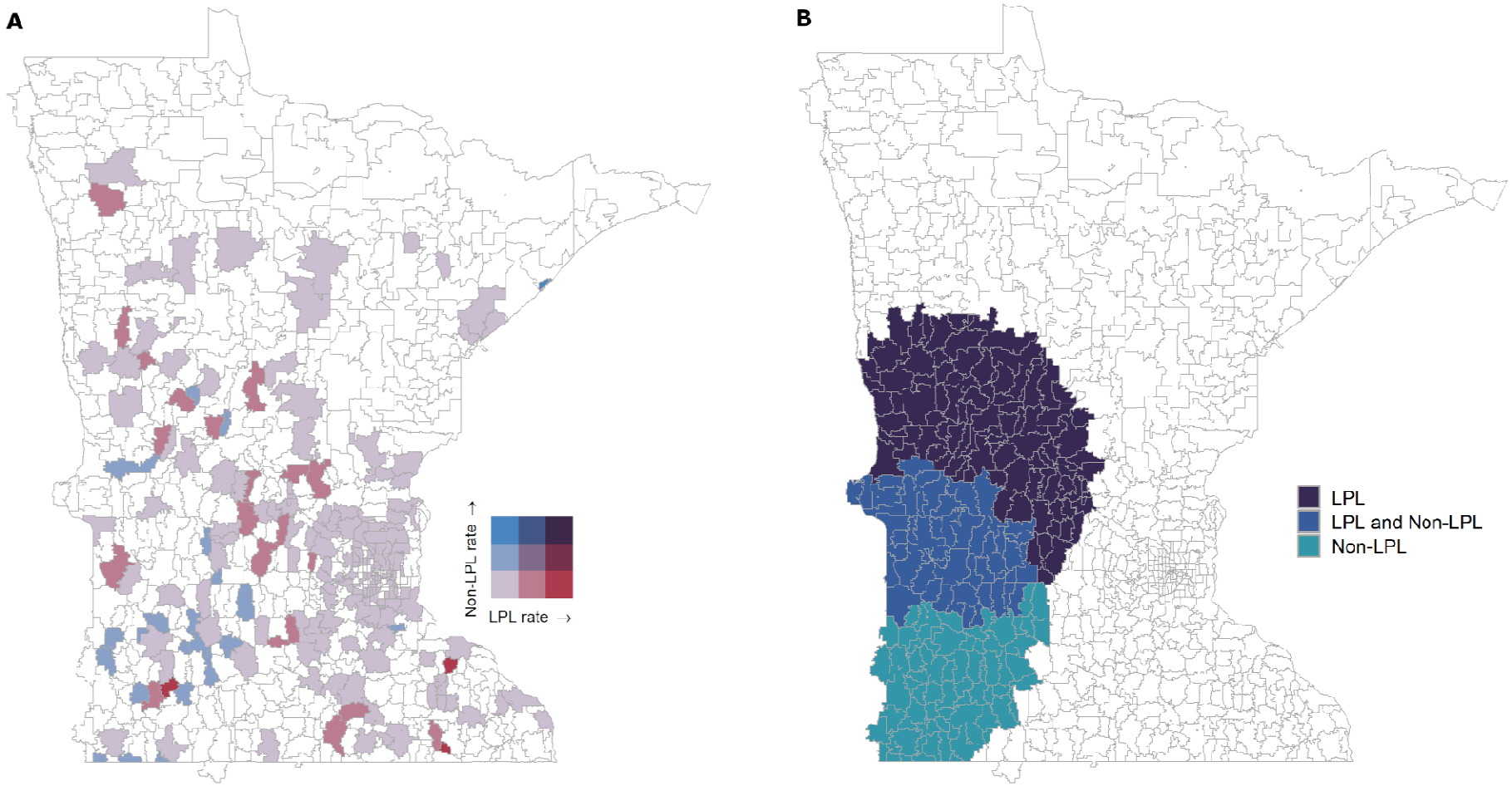


Figure 3. Spatial distribution of LPL and non-LPL isolates. Cases were aggregated to zip code tabulation areas (ZCTAs) based on home address. A) The incidence rate of LPL and non-LPL isolates appeared to differ in some parts of the state and overlap in others. B) Poisson-based spatial scan statistics limited to 10% of the population identified individual clusters of LPL and non-LPL isolates, which partially overlapped. The incidence of an LPL-associated STEC O157:H7 infection was 2.97 times greater ($p = 0.001$) within the cluster relative to the risk in the entire state. The incidence of a non-LPL-associated STEC O157:H7 infection was 3.62 times greater ($p = 0.001$) within the cluster relative to the risk in the entire state.

642 # Tables

643 Table 1. Characteristics of analyzed isolates and Minnesota LPLs, by *E. coli* O157:H7 clades

Characteristic	All O157	B	C	D	E	F	G	Difference by clade, p-value
Isolates								
MN isolates	363	0	8	6	78	151	120	
External isolates	1,169	4	7	20	84	661	338	
LPLs								
LPLs, n	15	0	1	1	3	3	7	
MN isolates, n (%; 95% CI)	128 (35.3%; 30.3%, 40.4%)	-	3 (37.5%; 8.5%, 75.5%)	3 (50.0%; 11.8%, 88.2%)	20 (25.6%; 16.4%, 36.8%)	45 (29.8%; 22.6%, 37.8%)	57 (47.5%; 38.3%, 56.8%)	0.006 [a]
External isolates, n (%; 95% CI)	21 (1.8%; 1.1%, 2.7%)	-	1 (14.3%; 0.4%, 57.9%)	0 (0%; 0%, 16.8%)	4 (4.8%; 1.3%, 11.7%)	8 (1.2%; 0.5%, 2.4%)	8 (2.1%; 0.9%, 4.0%)	0.038 [a]
Isolates per LPL, mean (SD) [b]	9.9 (6.8)	-	4	3	8.0 (4.4)	17.7 (5.9)	9.3 (6.6)	0.137 [c]
Maximum SNP distance, mean (SD) [b]	34 (21)	-	79	62	19 (9)	46 (23)	26 (8)	0.059 [c]
Ancestor year, mean (SD) [b]	2004 (9)	-	1998	1979	2009 (6)	2002 (4)	2007 (6)	0.363 [c]
Duration (years), mean (SD) [b]	3.8 (2.8)	-	2.1	2.23	2.5 (0.8)	8.1 (0.8)	3.0 (2.5)	0.008 [c,d]

644 [a] Tested using Fisher's exact test; clade B excluded.

645 [b] Clades C and D each have only 1 LPL, so value for that LPL is shown rather than a distribution.

646 [c] Tested using ANOVA; clades B, C, and D excluded.

647 [d] Residuals were non-normal due to a high outlier in clade G; sensitivity analysis of the difference between clades E, F, and G after removal of the outlier
648 yielded $p < 0.001$.

649

650

651

652 Table 2. Comparison of LPL and non-LPL isolates from Minnesota, 2010-2019

	LPL (n=128)	***Non-LPL (n=235)***	***Chi-square p-value***
Outbreak-associated, n (%; 95% CI)			<0.001
Minnesota outbreak	40 (31.3%; 23.4%, 40.0%)	12 (5.1%; 2.7%, 8.7%)	
Multi-state outbreak	0 (0%; 0%, 2.8%)	39 (16.6%; 12.1%, 22.0%)	
Sporadic	88 (68.8%; 60.0%, 76.6%)	184 (78.3%; 72.5%, 83.4%)	
MN outbreak type, n (%; 95% CI)			0.596 [a]
Animal contact	13 (32.5%; 18.6%, 49.1%)	3 (25.0%; 5.5%, 57.2%)	
Foodborne	17 (42.5%; 27.0%, 59.1%)	4 (33.3%; 9.9%, 65.1%)	
Person-to-person	10 (25.0%; 12.7%, 41.2%)	5 (41.7%; 15.2%, 72.3%)	
Clinical outcomes, n (%; 95% CI)			
Bloody diarrhea	85 (66.4%; 57.5%, 74.5%)	167 (71.1%; 64.8%, 76.8%)	0.759
Hospitalized	53 (41.4%; 32.8%, 50.4%)	106 (45.1%; 38.6%, 51.7%)	0.744
HUS	24 (18.8%; 12.4%, 26.6%)	51 (21.7%; 16.6%, 27.5%)	0.684
***stx* profile, n (%; 95% CI)**			<0.001 [a]
stx1a	1 (0.8%; 0%, 4.3%)	1 (0.4%; 0%, 2.3%)	
stx1a/stx2a	50 (39.1%; 30.6%, 48.1%)	31 (13.2%; 9.1%, 18.2%)	
stx1a/stx2c	24 (18.8%; 12.4%, 26.6%)	50 (21.3%; 16.2%, 27.1%)	
stx2a	8 (6.3%; 2.7%, 11.9%)	38 (16.2%; 11.7%, 21.5%)	
stx2c	2 (1.6%; 0.2%, 5.5%)	19 (8.1%; 4.9%, 12.3%)	
stx2a/stx2c	38 (29.7%; 21.9%, 38.4%)	90 (38.3%; 32.1%, 44.8%)	
Other *stx1/stx2*	5 (3.9%; 1.3%, 8.9%)	6 (2.6%; 0.9%, 5.5%)	
Other virulence genes, n (%; 95% CI) [b]			
ehxA	125 (97.7%; 93.3%, 99.5%)	221 (94.0%; 90.2%, 96.7%)	0.192 [a]
espP	122 (95.3%; 90.1%, 98.3%)	227 (96.6%; 93.4%, 98.5%)	0.575 [a]
ARGs present, n (%; 95% CI)	6 (4.7%; 1.7%, 9.9%)	42 (17.9%; 13.2%, 23.4%)	0.001
Stress tolerance genes, n (%; 95% CI) [c]			
emrE	110 (85.9%; 78.7%, 91.4%)	191 (81.3%; 75.7%, 86.1%)	0.326
qacE	0 (0%; 0%, 2.8%)	18 (7.7%; 4.6%, 11.8%)	0.003
Mercury tolerance [d]	4 (3.1%; 0.9%, 7.8%)	19 (8.1%; 4.9%, 12.3%)	0.104

653 Abbreviations: ARG, antimicrobial resistance gene; CI, confidence interval; HUS, hemolytic uremic syndrome; LPL,
654 locally persistent lineage

655 [a] Tested using Fisher's exact test.

656 [b] Virulence genes *eae*, *iha*, and *tir* were present in all isolates.

657 [c] Tolerance genes for arsenic (*arsC* and *arsR*) and tellurium (*terD*, *terW*, and *terZ*) were present in all isolates. Tolerance
658 genes for acid (*asr*) were present in 99.7% of isolates.

659 [d] Includes any combination of genes *merA*, *merB*, *merC*, *merD*, *merE*, *merP*, *merR*, and *merT*.

660