Geogenomic Segregation and Temporal Trends of Human Pathogenic *Escherichia coli* O157:H7, Washington, USA, 2005–2014[1]

Gillian A.M. Tarr, Smriti Shringi, Amanda I. Phipps, Thomas E. Besser, Jonathan Mayer, Hanna N. Oltean, Jon Wakefield, Phillip I. Tarr, Peter Rabinowitz

The often-noted and persistent increased incidence of *Escherichia coli* O157:H7 infections in rural areas is not well understood. We used a cohort of *E. coli* O157:H7 cases reported in Washington, USA, during 2005–2014, along with phylogenomic characterization of the infecting isolates, to identify geographic segregation of and temporal trends in specific phylogenetic lineages of *E. coli* O157:H7. Kernel estimation and generalized additive models demonstrated that pathogen lineages were spatially segregated during the period of analysis and identified a focus of segregation spanning multiple, predominantly rural, counties for each of the main clinical lineages, Ib, IIa, and IIb. These results suggest the existence of local reservoirs from which humans are infected. We also noted a secular increase in the proportion of lineage IIa and IIb isolates. Spatial segregation by phylogenetic lineage offers the potential to identify local reservoirs and intervene to prevent continued transmission.

Escherichia coli O157:H7 infections cause major public health challenges. Most *E. coli* O157:H7 infections occur sporadically, and the source of infection is often difficult to identify with certainty (*1,2*). Many reported infections are attributed to food vehicles (*1*), but studies have implicated other risk factors, and environmental transmission may be particularly notable in rural areas (*3–7*). Overall, the frequency of infections with *E. coli* O157:H7 has fallen in the United States, which is likely related to improved food safety (*8*), but it is not clear that rural incidence has also fallen.

Residing in a rural area confers increased risk for *E. coli* O157:H7 infection (*9,10*). *E. coli* O157:H7 can persist in certain locales, posing ongoing risk to humans. Multiple studies demonstrate that specific strains persist within cattle farms and spread to neighboring farms (*11–15*). The reservoirs enabling this persistence may include water, soil, and wild birds (*16–19*). It is, therefore, possible that humans incidentally acquire *E. coli* O157:H7 infections because they reside in a geographic region with a persistent reservoir. Using a generalizable population-based cohort, we sought to test the hypothesis that there are geographic foci of related *E. coli* O157:H7 infections, most likely of environmental origin, taking into account the genomic relatedness of different isolates (*20,21*) and the geographic, temporal, and secular attributes of their corresponding infections.

Methods

Study Population and Pathogen Characterization

We conducted a population-based retrospective cohort study of all culture-confirmed *E. coli* O157:H7 cases reported to the Washington State Department of Health (DOH; Shoreline, WA, USA) during 2005–2014. *E. coli* O157:H7 case reporting mandated by the Washington Administrative Code occurs primarily through diagnostic laboratories and healthcare providers. Local health jurisdictions use a standardized DOH case report form to abstract medical records; interview case-patients to obtain demographic information (including residence address), potential exposures, and details of the course of illness; and determine the most likely source of infection. For this study, case addresses were geocoded and census block groups determined. Case data were deidentified for analysis. This study was deemed exempt by the Washington State Institutional Review Board.

All *E. coli* O157:H7 isolates are sent to DOH for microbiologic confirmation and XbaI pulsed-field gel electrophoresis (PFGE) typing. We obtained isolates from

Author affiliations: University of Calgary, Calgary, Alberta, Canada (G.A.M. Tarr); Washington State University, Pullman, Washington, USA (S. Shringi, T.E. Besser); University of Washington, Seattle, Washington, USA (A.I. Phipps, J. Mayer, J. Wakefield, P. Rabinowitz); Washington State Department of Health, Shoreline, Washington, USA (H.N. Oltean); Washington University School of Medicine, St. Louis, Missouri, USA (P.I. Tarr)

DOI: https://doi.org/10.3201/eid2401.170851

[1]Preliminary results from this study were presented at the International Meeting on Emerging Diseases and Surveillance (IMED), November 4–7, 2016, Vienna, Austria.

DOH and determined their lineage according to the phylogenetic tree developed by Bono et al. (*20*) and expanded by Jung et al., who identified some lineages as clinical and others as bovine-biased (*21*). We used the Jung et al. 48-plex single-nucleotide polymorphism (SNP) assay to type a subset of isolates (*21*). We assumed that all isolates with a given PFGE pattern would be SNP typed to the same lineage. Thus, we typed ≥1 isolate from each PFGE pattern in the dataset and inferred the lineage of nontyped isolates. Concordance among isolates with identical PFGE profiles was confirmed (online Technical Appendix, https://wwwnc.cdc.gov/EID/article/24/1/17-0851-Techapp1.pdf). We analyzed the clinically common lineages Ib, IIa, and IIb separately and analyzed the bovine-biased and remaining sparsely represented lineages (*21*) as a clinically rare group.

Phylogenetic Lineage Spatial Segregation

Spatial segregation is the ecologic concept that one species or species type is more likely to be surrounded by like than by nonlike individuals (*22*). We used Diggle's kernel estimation method (*23*) and spatialkernel package (*24*) in R (*25*) to test spatial segregation of *E. coli* O157:H7 by phylogenetic lineage (online Technical Appendix). In brief, we first estimated a smoothed probability surface for each lineage by comparing the distance between cases infected with the same lineage to the distance between cases infected with different lineages. A peak in the lineage-specific probability surface indicates an area with a high probability of that lineage, relative to the distribution of the other lineages. For example, if 80% of cases in a given proximity are infected with lineage Ib but in all other areas lineage Ib causes only 50% of cases, we would observe a peak in the lineage Ib-specific probability surface, suggesting spatial segregation. To determine overall spatial segregation, the probability surfaces were compared with a null distribution in which the proportion of infections caused by each lineage is constant across space.

We next sought to account for potential confounders and to detect geographic trends. To do so, we modeled the risk surface using a multinomial generalized additive model (GAM). We estimated the effect of a bivariate thin plate regression spline smooth of latitude and longitude on the odds of infection with a given lineage compared with the most common lineage. This smoothing technique produces a risk surface that can vary flexibly across both horizontal and vertical coordinates. In this analysis, we compared lineages IIa and IIb and the group of clinically rare lineages separately with lineage Ib, which served as the reference (most common) lineage. The model was adjusted for sex and age group (<5, 5–9, 10–19, 20–59, and ≥60 years); isolates from cases of unknown age (n = 1) or sex (n = 10) were excluded from analysis. We estimated parameters using restricted maximum likelihood and the mgcv package in R (*26,27*). We further conducted a series of sensitivity analyses to determine the robustness of our results by seeking to confirm our results with 2 independent methods: Dixon's nearest-neighbor test (*22*) and multinomial spatial scan statistics (*28*) (online Technical Appendix).

Temporal Variation in Spatial Segregation

To determine whether spatial segregation of lineages varied over time, we replicated our spatial segregation analyses incorporating time. To do so, we split the years of analysis into 3 intervals (2005–2007, 2008–2010, and 2011–2014) and calculated a kernel-based estimate of spatial segregation for each. We evaluated the effect of time in the multinomial GAM by adding year to the model as a continuous variable, testing the effect of year as both a linear term and as a smoothed term using a thin plate spline. The thin plate spline allows the association between lineage and year to smoothly change in magnitude and direction.

Exploratory Risk Factor Analysis

We explored potential drivers of segregation by testing the association of risk factors included on the DOH case report form with each lineage compared with the reference lineage Ib. Using multinomial GAMs adjusting for sex, age, year, and latitude and longitude as a thin plate spline bivariate smoother, we tested each risk factor (online Technical Appendix Table 1). In addition to the statewide analyses, region-specific analyses were conducted for the 3 regions with the highest *E. coli* O157:H7 incidence to determine locally key associations. Regions were defined based on major population centers, areas of increased agricultural intensity, and observed segregation clusters, and models were adjusted for sex, age, and year.

Results

During the study period, 1,160 *E. coli* O157:H7 cases were reported to DOH. Of these, 33 isolates, representing 31 PFGE types, were not available for typing (online Technical Appendix), and isolates from 6 cases were excluded as biochemically atypical *E. coli* O157:H7 (online Technical Appendix Figure 1). We SNP typed 793 isolates and, by extension, matched another 328 to a known lineage using PFGE, enabling us to assign a specific lineage of *E. coli* O157:H7 to isolates from 1,121 cases. Ten cases lacked address data and were excluded, leaving 1,111 cases for analysis.

Lineages Ib, IIa, and IIb, in descending order, were the most common lineages (Table). Twelve clinically rare lineages were identified, including 2 not previously described, encompassing 45 unique PFGE types (online Technical Appendix Figure 1). Lineage Ib comprised 210 PFGE types, whereas lineage IIa comprised only 38 PFGE types

Table. *Escherichia coli* O157:H7 lineage frequency by case characteristic among culture-confirmed human cases reported in Washington, USA, 2005–2014*

Variable	Lineage Ib	Lineage IIa	Lineage IIb	Rare lineages†
Total	586 (52.7)	260 (23.4)	199 (17.9)	66 (5.9)
Mean isolates per PFGE type (SD)‡	2.8 (5.3)	6.8 (14.3)	7.7 (24.7)	1.5 (1.7)
Sex				
F	333 (56.8)	163 (62.7)	105 (52.8)	33 (50.0)
M	244 (41.6)	97 (37.3)	94 (47.2)	32 (48.5)
Unknown	9 (1.5)	0	0	1 (1.5)
Age group, y				
<5	119 (20.3)	72 (27.7)	63 (31.7)	10 (15.2)
5–9	81 (13.8)	32 (12.3)	33 (16.6)	12 (18.2)
10–19	97 (16.6)	51 (19.6)	31 (15.6)	6 (9.1)
20–59	207 (35.3)	81 (31.2)	49 (24.6)	29 (43.9)
≥60	81 (13.8)	24 (9.2)	23 (11.6)	9 (13.6)
Unknown	1 (0.2)	0	0	0
HUS				
Yes	37 (6.3)	18 (6.9)	20 (10.0)	0
No	526 (89.2)	236 (90.1)	173 (86.1)	67 (98.5)
Unknown	27 (4.6)	8 (3.1)	8 (4.0)	1 (1.5)

*Values are no. (%) except as indicated. HUS, hemolytic uremic syndrome; PFGE, pulsed-field gel electrophoresis.
†Twelve clinically rare lineages.
‡PFGE type percentages indicate the proportion of PFGE types with an assigned lineage (n = 355) belonging to each lineage.

and lineage IIb 26 PFGE types (online Technical Appendix Figure 1). Lineage IIa contained an average of 7 (SD 14) and IIb an average of 8 (SD 25) isolates per PFGE type, compared with 3 (SD 5) for lineage Ib and 1 (SD 2) for the clinically rare lineages (Table).

Distribution of cases by sex, age group, and hemolytic uremic syndrome (HUS) status varied by lineage (Table). Lineage IIa and IIb isolates originated disproportionately from children <5 years of age compared with isolates in lineage Ib. Patients infected with lineage IIb bacteria also had higher frequencies of HUS (10%) than other patients (6%). None of the patients with infections caused by isolates from the clinically rare lineages developed HUS.

Spatial Segregation

The result of Diggle's kernel estimation test was statistically significant (p = 0.001), suggesting spatial segregation. Lineage-specific probability surfaces showed separate, distinct peaks for lineages Ib, IIa, and IIb (Figure 1). The southwest region of Washington was marked by segregation of lineage IIb isolates and correspondingly lower probability of isolating lineage Ib from cases. Spatial segregation was observed for lineage Ib isolates in northwest Washington and for lineage IIa isolates in the south-central region. There was low probability of lineage IIb isolates in both these areas. Sensitivity analysis corroborated these results (online Technical Appendix).

Consistent with the kernel regression results, the adjusted GAM risk surface of lineage IIb varied significantly from that of Ib (p<0.001), providing additional support of the spatial segregation. The frequency of lineage IIb isolation was greater than the frequency of Ib in the southwest region, but this imbalance diminished as latitude and longitude increased (Figure 2), that is, in areas northward and eastward. This spatial pattern was also observed in the kernel estimation map of lineage IIb (Figure 1). The risk surfaces of lineage IIa and the clinically rare lineage group did not differ significantly from that of Ib (online Technical Appendix Table 2). In sensitivity analyses designed to gauge the robustness of results to model assumptions, the spatial risk surface of lineage IIb consistently varied significantly from the risk surface of lineage Ib (online Technical Appendix Table 2). The spatial risk surface of lineage IIa also varied significantly from the risk surface of lineage Ib in some sensitivity analyses, similar to the spatial distribution in the kernel estimation lineage IIa–specific probability surface.

We also found significant differences in lineage by age of infected patients, independent of geography. The likelihood of being an adult (age ranges 20–59 and ≥60 years of age) versus being a toddler (<5 years of age) was lower among IIa-infected patients than among Ib-infected patients (20–59 years odds ratio [OR] 0.65, 95% CI 0.44–0.96; ≥60 years OR 0.49, 95% CI 0.28–0.85). The odds of being 20–59 years of age versus <5 years were also lower among IIb-infected patients than among Ib-infected patients (OR 0.44, 95% CI 0.28–0.69). Thus, adults comprised a smaller proportion of patients infected with lineage IIa or IIb *E. coli* O157:H7 than of those infected with lineage Ib. We found no significant differences by sex.

Temporal Variation

The incidence of *E. coli* O157:H7 averaged 1.73/100,000 population during the study period. Although incidence fluctuated from a low of 1.37/100,000 population in 2014 to a maximum of 2.28/100,000 population in 2013, we found no discernible trend in overall incidence. However,

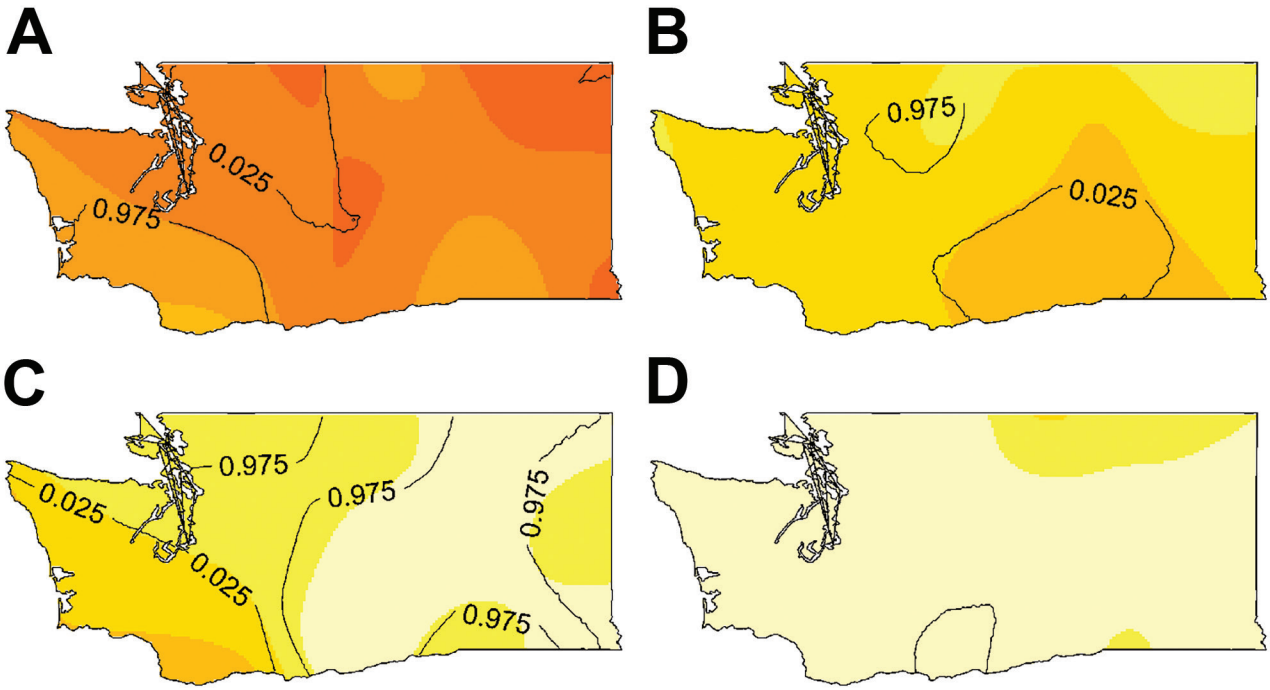


Figure 1. *Escherichia coli* O157:H7 lineage frequency among culture-confirmed human cases reported in Washington, USA, 2005–2014. A) Lineage Ib; B) lineage IIa; C) lineage IIb; D) rare lineages (12 different clinically rare lineages). Lineage-specific probability surfaces were determined by kernel-based estimation of spatial segregation. Darker shading indicates higher risk for that lineage. Contour lines marked 0.025 define areas in which there is a high probability of cases being caused by a given lineage, suggesting spatial segregation. Contour lines marked 0.975 define areas in which there is a low probability of cases being caused by the given lineage.

the composition of the *E. coli* O157:H7 population shifted over time (Figure 3). In the GAM analysis including year as a linear term, incidence relative to lineage Ib increased over time for lineage IIa (OR 1.26, 95% CI 1.19–1.34), lineage IIb (OR 1.10, 95% CI 1.03–1.17), and clinically rare lineages (OR 1.13, 95% CI 1.02–1.26).

We observed a peak of lineage IIb incidence during the middle of the study period in southwest Washington and the Seattle–Tacoma region (Figure 3). Using kernel regression, we identified statistically significant temporal variation in spatial segregation across intervals (p = 0.001). We observed statistically significant overall spatial segregation only during the 2008–2010 interval (p = 0.001). Some portion of the southwest region of the state showed increased probability of lineage IIb isolation during all intervals, and lineages Ib and IIa were segregated during 2008–2010 and 2011–2014 (Video, https://wwwnc.cdc.gov/EID/article/24/1/17-0851-V1.htm). Cross-validated log-likelihood bandwidths used in these analyses ranged from 0.73 to 1.0. In sensitivity analysis, a lower bandwidth yielded statistically significant spatial segregation during all periods (online Technical Appendix). Latitude and longitude remained significant predictors of Ib in GAMs that included year (online Technical Appendix Table 2).

Sensitivity Analysis

Alternate analytic approaches confirmed the results of our primary analyses. Dixon's test for spatial segregation identified statistically significant spatial segregation overall, as well as for lineages Ib, IIa, and IIb (online Technical Appendix Tables 3, 4). Three clusters identified using multinomial spatial scan statistics paralleled areas of segregation found in the kernel regression analysis and were consistent with the southwest trend toward proportionally greater IIb observed in the multinomial GAM (online Technical Appendix Figures 3, 4).

To focus on potential local reservoirs, which are not likely to be human, we also conducted the analysis without cases due to presumptive person-to-person transmission (online Technical Appendix). We used the most likely source of infection documented on the DOH case report form to exclude patients most likely infected by other persons. After discounting secondary transmission, we observed spatial segregation using the kernel estimation method (p = 0.002). The risk surface of lineage IIb still varied significantly from that of Ib (p<0.001). The trend toward greater IIb relative to Ib risk in southwest Washington was consistent with the analysis of all cases, but relative IIb risk was substantially lower in the northeast region than that observed in the primary analysis.

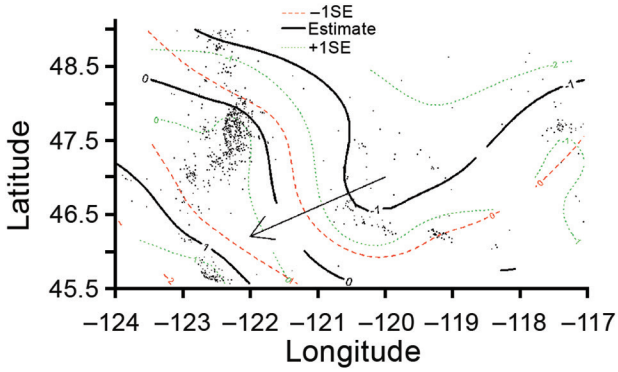


Figure 2. Risk surface of *Escherichia coli* O157:H7 lineage IIb relative to lineage Ib using a multinomial generalized additive model and a bivariate thin plate smooth function for longitude and latitude for culture-confirmed human cases reported in Washington, USA, 2005–2014. The black contour lines show the mean effect estimate for lineage IIb relative to Ib as latitude and longitude change. The 0-marked black line indicates no effect. The 1-marked black line indicates greater proportional incidence of lineage IIb toward the southwest corner of the area as compared to lineage Ib (p<0.001). The arrow indicates the general direction of the trend from higher Ib risk to higher IIb risk. Dashed red lines show the effect estimate 1 standard error (SE) below (to the south and west) the mean estimate. Dotted green lines show the effect estimate 1 standard error above (to the north and east) the mean estimate.

This pattern suggests that lineage IIb infections in northeast, but not southwest, Washington may be disproportionately attributed to secondary transmission compared with Ib infections. Finally, we found no evidence of case ascertainment bias that could independently explain our results (online Technical Appendix).

Exploratory Risk Factor Analysis

Statewide, patients infected with lineage IIa *E. coli* O157:H7 were more likely to have reported raw fruit or vegetable consumption than those infected with lineage Ib pathogens (OR 1.81, 95% CI 1.05–3.11). Patients infected with lineage IIb *E. coli* O157:H7 were more likely to have reported raw milk consumption than those infected with lineage Ib pathogens (OR 2.46, 95% CI 1.15–5.28). All examined risk factors and associations are summarized in online Technical Appendix Table 1.

Discussion

The geographic differences and temporal trends in the relative frequencies of lineages of *E. coli* O157:H7 from cases in Washington demonstrate that, in addition to genomic variation reported at the national level (*29,30*), persistent geogenomic variation exists at the regional level. Several geospatial associations warrant elaboration. In all analyses, lineage IIb cases were segregated in the southwest region of the state. Southwest Washington includes Olympia, the state capital, as well as suburbs of Portland, Oregon, north of the Columbia River; however 27% of the population in the 12 southwest region counties is considered rural, compared with 16% of the state as a whole (*31*). Small farms are common. The southwest region is home to >20% of the state's farms but accounts for only 7.1% of its cattle and 6.3% of farm acreage (*32*). Roosevelt elk roam the southwest region, and elk elsewhere in the country have been identified as Shiga toxin–producing *E. coli* carriers (*33*). Water is also a potential factor in *E. coli* O157:H7 epidemiology in the southwest region, which has abundant coastal and river exposures. The largest recognized IIb outbreak in this region accounted for only 11 cases linked to a particular daycare center (out of 77 IIb infections in the region), so the observed segregation is unlikely due to a single point source. Notably, lineages IIa and IIb have the greatest overlap with the putatively hypervirulent clade 8 (*34*), making their segregation of particular concern.

Lineage IIb isolates were relatively uncommon in the northwest and south-central regions of Washington, both major cattle-production regions. Lineage Ib showed segregation in the northwest and IIa in the south-central region in some analyses, although their adjusted risk surfaces did not differ significantly, suggesting overlap. More research is needed to clarify why lineage IIb has not yet also established itself in areas with abundant cattle.

The presence of spatially segregated lineages indicates local environmental reservoirs producing infections above and beyond those caused by widely distributed exogenous sources such as food. We propose that persistent spatial segregation of a lineage could reflect a founder effect, in which an ancestral pathogen has become established in a region, persisted, and expanded and occasionally crosses into the human population. Such a dynamic would result in phylogenetically similar bacteria being isolated in the same general geographic region separated by months or years, as we have observed in this study. A possible precedent exists in a report of 2 cases from Webster County, Missouri, USA (*35*). Our findings are also consistent with those of Jaros et al., who found that geography explains some variation in *E. coli* O157:H7 strains in New Zealand (*36*). In addition, prior work from Washington demonstrated shifts over time in the Shiga toxin genotypes of *E. coli* O157:H7 (*37*).

The clinical infections in our study were dominated by *E. coli* O157:H7 in lineages Ib, IIa, and IIb, consistent with the results of Jung et al. (*21*). Our work is also consistent with a national study showing that lineage Ib *E. coli* O157:H7 causes most clinical cases in the United States (*30*). Relative to lineage Ib, Washington experienced statistically significant increases in the other clinically common lineages during the study period. The increase is most dramatic for lineage IIa, which appears to have emerged in most regions in the latter half of the study period (Figure 3).

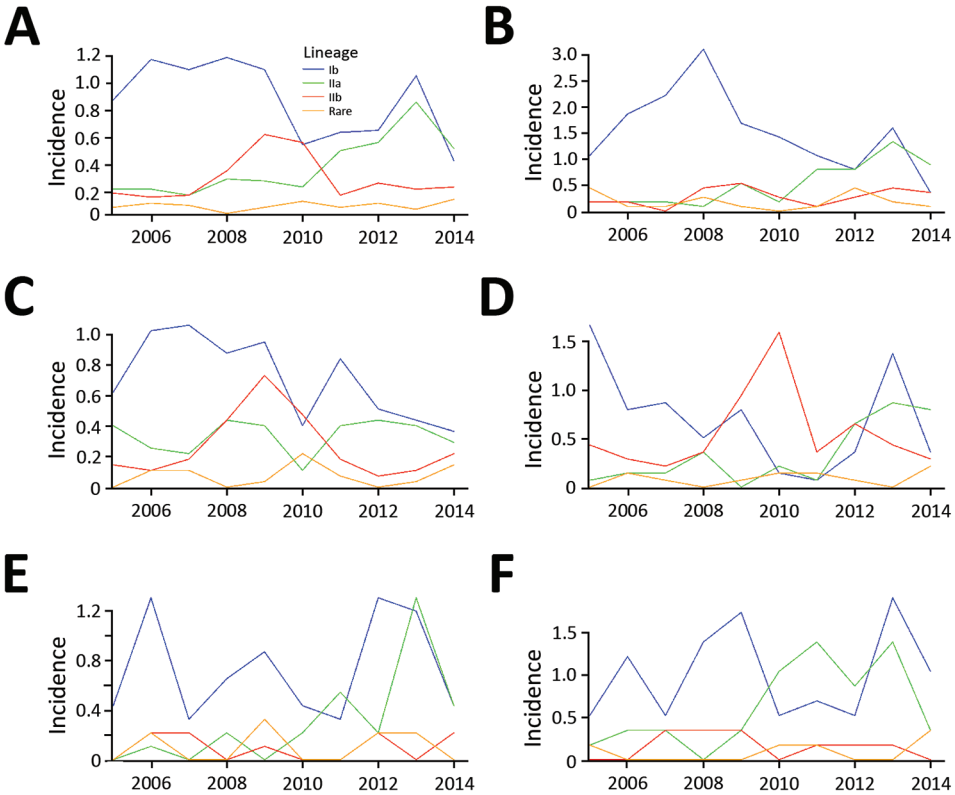


Figure 3. Annual incidence (per 100,000 population) of reported *Escherichia coli* O157:H7 cases by phylogenetic lineage, Washington, USA, 2005–2014. A) Statewide; B) northwest region; C) Seattle–Tacoma region; D) southwest region; E) northeast region; F) south-central region. Regions were defined according to major demographic characteristics and patterns of segregation observed in analyses for the whole period. The northwest region experienced the highest peak incidence. The Seattle–Tacoma region and the northeast region experienced the lowest incidences. "Rare" indicates 12 different clinically rare lineages. **Video.** Lineage-specific probability surfaces for *Escherichia coli* O157:H7 from culture-confirmed human cases reported in Washington, USA, 2005–2014. Probabilities were determined by kernel-based estimation of spatial segregation for 3 intervals: 2005–2007 (n = 305, bandwidth = 1.0000); 2008–2010 (n = 367, bandwidth = 0.7256); and 2011–2014 (n = 439, bandwidth = 0.9314). Overall spatial segregation was not statistically significant for the 2005–2007 interval (p = 0.769) or 2011–2014 interval (p = 0.138) but was statistically significant for the 2008–2010 interval (p = 0.001). Circles indicate case locations. Darker hues indicate higher risk. Contour lines marked 0.025 define areas in which there is a high probability of cases being caused by a given lineage, suggesting spatial segregation. There is an area of statistically significant spatial segregation for lineage IIb in all 3 intervals. Contour lines marked 0.975 define areas in which there is a low probability of cases being caused by the given lineage.

This difference could reflect the changing epidemiology of *E. coli* O157:H7 discussed by Rivas et al., owing to changes in food sources and consumption, or, possibly, pathogen evolution (*38*). Lineage IIa *E. coli* O157:H7 has emerged as a major cause of disease across the state, suggesting a disseminated driver of infections for this lineage overall. Lineage IIa's observed association with raw fruit and vegetable consumption, as compared with that for lineage Ib, is consistent with this hypothesis. The south-central region of Washington, identified in some analyses as an area of IIa segregation, experienced an uptick in IIa infections earlier than in other regions. This area includes the Yakima Valley, an area of higher agricultural intensity; a local IIa reservoir in this region could produce the observed segregation independent of statewide trends.

Our findings suggest exposures that may be preferentially associated with particular lineages. Specifically, we observed associations of lineage IIb with drinking untreated/unchlorinated water and raw milk in the southwest region, where this lineage is segregated (online Technical Appendix Table 1). There may be a lineage IIb reservoir in animals producing raw milk in this area, or bacteria from environmental reservoirs in the area may spill over into these animals and local water sources. Only 1 small, recognized raw milk outbreak in 2005 was noted on the DOH case report forms, making it unlikely that a single source is responsible for the association we found over time. It is possible that some *E. coli* O157:H7 lineages may be especially successful in surviving in particular vehicles or environments, such as raw produce or unpasteurized milk or water. Secular changes might also be the result of shifting environmental exposure risk if, for example, contact between a reservoir and humans varies over time. Better knowledge of small-intermediate area transmission patterns will open opportunities for intervention if reservoirs can be identified.

Our study is limited by its reliance on SNP data to define phylogenetic lineages. Whole-genome sequencing would have supported finer resolution of relatedness, particularly among isolates that were segregated in time and space, and enabled us to trace the history of segregated clusters. Such an analysis would not necessarily alter our conclusions, however, because evolution of specific clades of *E. coli* O157:H7 within a region, and the identification of different sublineages, would still be consistent with a founder effect. In fact, the precise delineation of the chromosomal architecture in these pathogens might actually confirm a common progenitor, as demonstrated from worldwide analyses of *E. coli* O157:H7 (*39*). Our use of phylogenetic lineages rather than PFGE profiles is also a strength of the work, because PFGE does not put differences into evolutionary perspective (*39*). By basing the analysis on phylogenetic lineages, we captured relatedness among strains and indicate the level of *E. coli* O157:H7 diversity as it circulates through its host populations. We also used multiple analytic techniques to provide confidence that our results were not due to assumptions made by any particular method.

In summary, clusters of spatial segregation by phylogenetic lineage in Washington suggest local reservoirs that perennially cause human disease. Further exploration of land use, human movements, and social–behavioral factors could elucidate within-region drivers of spatial segregation. We see comparison of lineage-specific spatial patterns with distributions of these and other factors as an essential next step in understanding *E. coli* O157:H7 spatial segregation. Environmental risk assessment and longitudinal studies based on our findings would also provide valuable information by identifying pathogen reservoirs that have not been identified by traditional public health surveillance and that could be mitigated by public health or environmental measures. The makeup of the *E. coli* O157:H7 population in the state is also shifting. To manage emerging lineages, attention is needed to the heterogeneity in risk factors across the phylogenetic tree. Greater knowledge of the most likely sources of infection for particular lineages has the potential to focus both outbreak investigations and efforts to identify persistent reservoirs.

This work was supported by the National Institute of Environmental Health Sciences of the National Institutes of Health (award no. T32ES015459) and the National Institute of Allergy and Infectious Disease of the National Institutes of Health (award no. F31AI126834).

Dr. Tarr is a postdoctoral fellow in pediatric enteric infections at the University of Calgary, Calgary, Alberta, Canada. Her primary research interest is the maintenance, distribution, and virulence of zoonotic diseases affecting children.

References

1. Scallan E, Hoekstra RM, Angulo FJ, Tauxe RV, Widdowson MA, Roy SL, et al. Foodborne illness acquired in the United States—major pathogens. Emerg Infect Dis. 2011;17:7–15. http://dx.doi.org/10.3201/eid1701.P11101
2. Centers for Disease Control and Prevention (CDC). Vital signs: incidence and trends of infection with pathogens transmitted commonly through food—foodborne diseases active surveillance network, 10 U.S. sites, 1996–2010. MMWR Morb Mortal Wkly Rep. 2011;60:749–55.
3. Strachan NJ, Dunn GM, Locking ME, Reid TM, Ogden ID. *Escherichia coli* O157: burger bug or environmental pathogen? Int J Food Microbiol. 2006;112:129–37. http://dx.doi.org/10.1016/j.ijfoodmicro.2006.06.021
4. Denno DM, Keene WE, Hutter CM, Koepsell JK, Patnode M, Flodin-Hursh D, et al. Tri-county comprehensive assessment of risk factors for sporadic reportable bacterial enteric infection in children. J Infect Dis. 2009;199:467–76. http://dx.doi.org/10.1086/596555
5. Luffman I, Tran L. Risk factors for *E. coli* O157 and cryptosporidiosis infection in individuals in the karst valleys of east Tennessee, USA. Geosciences (Basel). 2014;4:202–18. http://dx.doi.org/10.3390/geosciences4030202
6. Michel P, Wilson JB, Martin SW, Clarke RC, McEwen SA, Gyles CL. Temporal and geographical distributions of reported cases of *Escherichia coli* O157:H7 infection in Ontario. Epidemiol Infect. 1999;122:193–200. http://dx.doi.org/10.1017/S0950268899002083
7. Locking ME, O'Brien SJ, Reilly WJ, Wright EM, Campbell DM, Coia JE, et al. Risk factors for sporadic cases of *Escherichia coli* O157 infection: the importance of contact with animal excreta. Epidemiol Infect. 2001;127:215–20. http://dx.doi.org/10.1017/S0950268801006045
8. Crim SM, Griffin PM, Tauxe R, Marder EP, Gilliss D, Cronquist AB, et al.; Centers for Disease Control and Prevention (CDC). Preliminary incidence and trends of infection with pathogens transmitted commonly through food—Foodborne Diseases Active Surveillance Network, 10 U.S. sites, 2006–2014. MMWR Morb Mortal Wkly Rep. 2015;64:495–9.
9. Haack JP, Jelacic S, Besser TE, Weinberger E, Kirk DJ, McKee GL, et al. *Escherichia coli* O157 exposure in Wyoming and Seattle: serologic evidence of rural risk. Emerg Infect Dis. 2003;9:1226–31. http://dx.doi.org/10.3201/eid0910.020254
10. Innocent GT, Mellor DJ, McEwen SA, Reilly WJ, Smallwood J, Locking ME, et al.; Wellcome Trust-funded IPRAVE Consortium. Spatial and temporal epidemiology of sporadic human cases of *Escherichia coli* O157 in Scotland, 1996–1999. Epidemiol Infect. 2005;133:1033–41. http://dx.doi.org/10.1017/S0950268805003687
11. Liebana E, Smith RP, Batchelor M, McLaren I, Cassar C, Clifton-Hadley FA, et al. Persistence of *Escherichia coli* O157 isolates on bovine farms in England and Wales. J Clin Microbiol. 2005;43:898–902. http://dx.doi.org/10.1128/JCM.43.2.898-902.2005
12. LeJeune JT, Besser TE, Rice DH, Berg JL, Stilborn RP, Hancock DD. Longitudinal study of fecal shedding of *Escherichia coli* O157:H7 in feedlot cattle: predominance and persistence of specific clonal types despite massive cattle population turnover. Appl Environ Microbiol. 2004;70:377–84. http://dx.doi.org/10.1128/AEM.70.1.377-384.2004
13. Rosales-Castillo JA, Vázquez-Garcidueñas MS, Alvarez-Hernández H, Chassin-Noria O, Varela-Murillo AI, Zavala-Páramo MG, et al. Genetic diversity and population structure of *Escherichia coli* from neighboring small-scale dairy farms. J Microbiol. 2011;49:693–702. http://dx.doi.org/10.1007/s12275-011-0461-2

14. Herbert LJ, Vali L, Hoyle DV, Innocent G, McKendrick IJ, Pearce MC, et al. *E. coli* O157 on Scottish cattle farms: evidence of local spread and persistence using repeat cross-sectional data. BMC Vet Res. 2014;10:95. http://dx.doi.org/10.1186/1746-6148-10-95
15. Widgren S, Söderlund R, Eriksson E, Fasth C, Aspan A, Emanuelson U, et al. Longitudinal observational study over 38 months of verotoxigenic *Escherichia coli* O157:H7 status in 126 cattle herds. Prev Vet Med. 2015;121:343–52. http://dx.doi.org/10.1016/j.prevetmed.2015.08.010
16. Saxena T, Kaushik P, Krishna Mohan M. Prevalence of *E. coli* O157:H7 in water sources: an overview on associated diseases, outbreaks and detection methods. Diagn Microbiol Infect Dis. 2015;82:249–64. http://dx.doi.org/10.1016/j.diagmicrobio.2015.03.015
17. Cernicchiaro N, Pearl DL, McEwen SA, Harpster L, Homan HJ, Linz GM, et al. Association of wild bird density and farm management factors with the prevalence of *E. coli* O157 in dairy herds in Ohio (2007–2009). Zoonoses Public Health. 2012;59:320–9. http://dx.doi.org/10.1111/j.1863-2378.2012.01457.x
18. Barker J, Humphrey TJ, Brown MWR. Survival of *Escherichia coli* O157 in a soil protozoan: implications for disease. FEMS Microbiol Lett. 1999;173:291–5. http://dx.doi.org/10.1111/j.1574-6968.1999.tb13516.x
19. Gargiulo A, Russo TP, Schettini R, Mallardo K, Calabria M, Menna LF, et al. Occurrence of enteropathogenic bacteria in urban pigeons (*Columba livia*) in Italy. Vector Borne Zoonotic Dis. 2014;14:251–5. http://dx.doi.org/10.1089/vbz.2011.0943
20. Bono JL, Smith TP, Keen JE, Harhay GP, McDaneld TG, Mandrell RE, et al. Phylogeny of Shiga toxin-producing *Escherichia coli* O157 isolated from cattle and clinically ill humans. Mol Biol Evol. 2012;29:2047–62. http://dx.doi.org/10.1093/molbev/mss072
21. Jung WK, Bono JL, Clawson ML, Leopold SR, Shringi S, Besser TE. Lineage and genogroup-defining single nucleotide polymorphisms of *Escherichia coli* O157:H7. Appl Environ Microbiol. 2013;79:7036–41. http://dx.doi.org/10.1128/AEM.02173-13
22. Dixon PM. Nearest-neighbor contingency table analysis of spatial segregation for several species. Écoscience. 2002;9:142–51. http://dx.doi.org/10.1080/11956860.2002.11682700
23. Diggle PJ, Zheng P, Durr P. Nonparametric estimation of spatial segregation in a multivariate point process: bovine tuberculosis in Cornwall, UK. Appl Stat. 2005;54:645–58. http://dx.doi.org/10.1111/j.1467-9876.2005.05373.x
24. Zheng P, Diggle PJ. Spatialkernel: nonparametric estimation of spatial segregation in a multivariate point process; R package version 0.4-19. 2013 [cited 2015 May 27]. https://CRAN.R-project.org/package=spatialkernel
25. R Core Team. R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2015.
26. Wood SN. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. J R Stat Soc Series B Stat Methodol. 2011;73:3–36. http://dx.doi.org/10.1111/j.1467-9868.2010.00749.x
27. Wood SN. Thin-plate regression splines. J R Stat Soc Series B Stat Methodol. 2003;65:95–114. http://dx.doi.org/10.1111/1467-9868.00374
28. Jung I, Kulldorff M, Richard OJ. A spatial scan statistic for multinomial data. Stat Med. 2010;29:1910–8. http://dx.doi.org/10.1002/sim.3951
29. Strachan NJ, Rotariu O, Lopes B, MacRae M, Fairley S, Laing C, et al. Whole genome sequencing demonstrates that geographic variation of *Escherichia coli* O157 genotypes dominates host association. Sci Rep. 2015;5:14145. http://dx.doi.org/10.1038/srep14145
30. Mellor GE, Fegan N, Gobius KS, Smith HV, Jennison AV, D'Astek BA, et al. Geographically distinct *Escherichia coli* O157 isolates differ by lineage, Shiga toxin genotype, and total Shiga toxin production. J Clin Microbiol. 2015;53:579–86. http://dx.doi.org/10.1128/JCM.01532-14
31. United States Census Bureau. 2010 Census urban and rural classification and urban area criteria. 2015 [cited 2015 May 27]. https://www.census.gov/geo/reference/ua/urban-rural-2010.html
32. United States Department of Agriculture. 2012 census of agriculture. Washington: National Agricultural Statistics Service; 2014.
33. Franklin AB, Vercauteren KC, Maguire H, Cichon MK, Fischer JW, Lavelle MJ, et al. Wild ungulates as disseminators of Shiga toxin-producing *Escherichia coli* in urban areas. PLoS One. 2013;8:e81512. http://dx.doi.org/10.1371/journal.pone.0081512
34. Manning SD, Motiwala AS, Springman AC, Qi W, Lacher DW, Ouellette LM, et al. Variation in virulence among clades of *Escherichia coli* O157:H7 associated with disease outbreaks. Proc Natl Acad Sci U S A. 2008;105:4868–73. http://dx.doi.org/10.1073/pnas.0710834105
35. Turabelidze G, Lawrence SJ, Gao H, Sodergren E, Weinstock GM, Abubucker S, et al. Precise dissection of an *Escherichia coli* O157:H7 outbreak by single nucleotide polymorphism analysis. J Clin Microbiol. 2013;51:3950–4. http://dx.doi.org/10.1128/JCM.01930-13
36. Jaros P, Cookson AL, Campbell DM, Duncan GE, Prattley D, Carter P, et al. Geographic divergence of bovine and human Shiga toxin–producing *Escherichia coli* O157:H7 genotypes, New Zealand. Emerg Infect Dis. 2014;20:1980–9. http://dx.doi.org/10.3201/eid2012.140281
37. Tarr PI, Neill MA, Clausen CR, Newland JW, Neill RJ, Moseley SL. Genotypic variation in pathogenic *Escherichia coli* O157:H7 isolated from patients in Washington, 1984–1987. J Infect Dis. 1989;159:344–7. http://dx.doi.org/10.1093/infdis/159.2.344
38. Rivas M, Chinen I, Miliwebsky E, Masana M. Risk factors for Shiga toxin–producing *Escherichia coli*–associated human diseases. Microbiol Spectr. 2014;2. http://dx.doi.org/10.1128/microbiolspec.EHEC-0002-2013
39. Leopold SR, Magrini V, Holt NJ, Shaikh N, Mardis ER, Cagno J, et al. A precise reconstruction of the emergence and constrained radiations of *Escherichia coli* O157 portrayed by backbone concatenomic analysis. Proc Natl Acad Sci U S A. 2009;106:8713–8. http://dx.doi.org/10.1073/pnas.0812949106

Address for correspondence: Gillian A.M. Tarr, Alberta Children's Hospital, Office C4-634, 2888 Shaganappi Trail NW, Calgary, AB T3B 6A8, Canada; email: gtarr@uw.edu

Article DOI: http://dx.doi.org/10.3201/eid2401.1710851

Geogenomic Segregation and Temporal Trends of Human Pathogenic *Escherichia coli* O157:H7

Technical Appendix 1

Supplementary Methods and Analyses

Assigning Phylogenetic Lineage to Non–SNP-Typed Isolates

In previous studies analyzing patterns associated with *E. coli* O157:H7 phylogenetic classification, it has been common to use a single representative isolate from each PFGE subtype (*1*–*3*). This practice masks the variability among isolates with the same PFGE fingerprint (e.g., variability in demographics, location). Further, estimation of effects at the population level is compromised, because the isolates being analyzed are not reflective of the *E. coli* O157:H7 case population distribution. To accurately make inference at the population level, we sought to include all reported cases during the study period. Because we did not have sufficient resources to SNP-type all isolates, we leveraged the assumption inherent in the single-representative-isolate approach, although not generally made explicit: isolates with the same PFGE fingerprint belong to the same phylogenetic grouping.

Our sample contained 1,160 isolates reflecting 355 unique XbaI PFGE patterns (Technical Appendix Figure 1). We SNP-typed 793 of these isolates, covering 319 PFGE subtypes. The 36 PFGE subtypes not SNP-typed were either biochemically atypical or they were not present in the isolate bank. Atypical isolates were exclusively from 2013 and 2014, the last 2 years of sampling. Missing isolates were predominantly (82%) from 2005 and 2006, the first 2 years of sampling. Of the 793 SNP-typed isolates, 570 belonged to a PFGE subtype with multiple SNP-typed isolates. Among these 570, we examined which phylogenetic lineages the isolates had been assigned via SNP typing. All but 1 PFGE subtype were assigned a consistent lineage. The one variable PFGE subtype was EXHX01.0047. It encompassed 82 isolates: 21 were not typed, 59 were typed to lineage IIa, and 2 were typed to lineage Ib. In other words, only

2 of 570 isolates (0.4%) showed aberrant lineage assignment. With this, we felt that the assumption that isolates of the same PFGE subtype would be in the same lineage held adequately well to use the SNP-typing results to assign lineage to non–SNP-typed isolates. We were able to assign lineage to 328 additional isolates by using this approach.

Spatial Segregation by Diggle's Kernel Estimation Method

Diggle's kernel estimation provides smoothed estimates of spatial segregation that take into account multiple neighbors of each case. It provides an overall test of spatial segregation and identifies statistically significant regions in the lineage-specific probability surfaces. Diggle's method assumes an underlying Poisson point process for each phylogenetic lineage. The degree of smoothing is dependent on the choice of a bandwidth. A cross-validated log-likelihood function can be used to calculate the bandwidth (*4*). We tested bandwidths between 0.02 and 1 degrees at 0.0098-degree increments to identify and then select for analysis the bandwidth (0.6472 degrees) associated with the greatest cross-validated log-likelihood. Using the selected bandwidth, we determined the lineage-specific probabilities based on the surrounding cases for each case location and plotted the lineage-specific probability surfaces on individual maps. We then calculated a test statistic for spatial segregation by summing the square of the difference between the kernel regression-estimated lineage-specific probability at a given location and the overall probability that a case isolate belongs to that lineage over all lineages and all case locations. To determine statistical significance, we performed 999 Monte Carlo replications with lineage randomly relabeled at each case location, maintaining the observed number of cases of each lineage. The proportion of test statistics greater than that observed from the data was the p-value. The analysis was conducted in R (*5*) using the spatialkernel package (*6*).

The bandwidth selected for the main analysis was used for all lineages within a given analysis. To identify the sensitivity of the kernel estimation results to the bandwidth of 0.6472 degrees that was selected, alternate bandwidths were tested: 0.02, 0.1, 0.2, 0.4, and 0.9. All yielded $p = 0.001$ for the overall test for spatial segregation. The segregation maps for individual lineages grew predictably smoother as the bandwidth was increased and identified statistically significant areas of segregation consistent with the primary result from a bandwidth of 0.6472.

Temporal variation in segregation was tested across 3 intervals: 2005–2007, 2008–2010, and 2011–2014. The slightly longer last interval is not expected to affect the validity of the

results. However, because of the greater number of cases in this interval, greater precision is expected. We calculated a new bandwidth for each new analysis and subset of the data using the cross-validated log-likelihood function. For the overall test of variation of spatial segregation across time intervals using the kernel regression method, we chose a bandwidth of 0.8236 degrees. The bandwidths chosen for each of the individual intervals were 1.0000 for 2005–2007, 0.7256 for 2008–2010, and 0.9314 for 2011–2014. Not unexpectedly, given the high degree of smoothing in the first and last periods, only the middle period had detectable overall spatial segregation ($p = 0.001$). However, all periods displayed some statistically significant spatial segregation for individual lineages (Technical Appendix Video). A bandwidth of 0.4 was also tested for each of the intervals, resulting in statistically significant tests for overall spatial segregation in each interval (2005–2007 $p = 0.037$, 2008–2010 $p = 0.001$, 2011–2014 $p = 0.014$).

Multinomial Generalized Additive Model

The multinomial GAM provides a smoothed risk surface relative to Ib, the most common lineage. Unlike the direct measures of spatial segregation, the GAM captures spatial trends without selecting a specific distance or number of neighbors across which to smooth. It does this through a flexible spline function. The GAM also supports adjustment for covariates, providing some assurance that the associations observed are not due to factors such as the distribution of cases by age. The multinomial analysis entailed logistic-type equations for each of the 3 lineage comparisons. Results of the GAM multinomial models must be interpreted conditional on having a reported *E. coli* O157:H7 illness. As such, odds ratios presented estimate risk proportional to that in the most common lineage, Ib.

We tested multiple aspects of the GAM specification. Latitude and longitude were specified individually and jointly to allow interaction. The basis dimension of the penalized regression smoother was altered to improve the effective degrees of freedom. Age and sex covariates were removed, and the form of the spline smoother was altered. Lineage IIa was used as the comparison lineage. These sensitivity analyses are summarized in Technical Appendix Table 2. None of the model perturbations meaningfully changed the primary model results. In the set of GAMs incorporating year, a trivariate smooth of latitude, longitude, and year was also tested and found to be statistically significant for lineages IIa and IIb (Technical Appendix Table 2).

Spatial Segregation by Dixon's Nearest-Neighbor Method

Another measure of spatial segregation, Dixon's nearest-neighbor method, considers only the closest neighbor of each case. It conducts no smoothing and can be expected to be sensitive to clustered outbreaks. This method does not indicate areas in which spatial segregation exists but does provide an overall test of spatial segregation, as well as for segregation of individual lineages and pairwise segregation tests. We created a 4 × 4 contingency table of nearest-neighbor counts for each lineage group. A χ^2 test with 12 degrees of freedom was used to test overall spatial segregation, and segregation was tested for each individual lineage group (Technical Appendix Table 3). We calculated Dixon's segregation index for each nearest-neighbor combination (e.g., from Ib to IIa; Technical Appendix Table 4). Dixon's pairwise segregation index is defined as:

$$S_{ij} = \log \frac{N_{ij}/(N_i - N_{ij})}{\mathrm{E}N_{ij}/(N_i - \mathrm{E}N_{ij})} = \log \frac{N_{ij}/(N_i - N_{ij})}{N_i/(N - N_j - 1)}$$

where *i* and *j* in this analysis are phylogenetic lineages (*7*). A positive value of *S* indicates association, and a negative value indicates segregation. We calculated Z-scores for each combination by comparing the observed nearest-neighbor count in each cell to the expected count. We calculated a p-value based on the Z-scores assuming an asymptotic normal distribution. We used the Dixon R package for this analysis (*8*).

We used Dixon χ^2 tests for segregation to indicate statistically significant segregation overall ($p < 0.001$) and for lineages Ib ($p = 0.046$), IIa ($p = 0.002$), and IIb ($p < 0.001$), but not for the group of clinically rare lineages (Technical Appendix Table 3). This is consistent with the findings of the kernel estimation method, which found statistically significant overall spatial segregation and identified areas of segregation for lineages Ib, IIa, and IIb. Dixon's method also tests associations between individual lineages. Pairwise nearest-neighbor comparisons showed statistically significant positive association from each of lineages Ib, IIa, and IIb to itself. Segregation was observed from Ib to IIa, IIa to the rare lineages, IIb to all other lineages, and the rare lineages to Ib (Technical Appendix Table 4).

We examined spatial segregation with Dixon's method for the 3 intervals analyzed with the kernel estimation method. Spatial segregation was found to be statistically significant with $p < 0.001$ during all 3 periods, contrasting with Diggle's method, which identified statistically

significant overall segregation only during the 2008–2010 period. However, the 2 spatial segregation tests were consistent in identifying spatial segregation of lineage IIb during all intervals ($p < 0.001$ for Dixon's method during all intervals). Additionally, Dixon's method identified segregation of lineage IIa during the 2005–2007 period ($p < 0.001$) and segregation of lineage Ib during the 2008–2010 ($p < 0.001$) and 2011–2014 ($p = 0.005$) periods.

Multinomial Spatial Scan Statistics

We used multinomial spatial scan statistics (*9*) in SaTScan (*10*) to identify clusters within which the distribution of lineages differed significantly from the distribution of lineages outside the cluster. The spatial scan statistics are designed to identify clusters of disease. In the multinomial framework used here, the clusters reflect areas within which the distribution of cases by lineage is skewed compared with the area outside the cluster. These are similar to the areas of segregation identified by the kernel regression method. However, the scan statistics look at the distribution of all 4 lineages simultaneously and not individually, thus allowing detection of clusters in which multiple lineages may be out of proportion. Like the multinomial GAM models, the multinomial spatial scan statistics must be interpreted conditionally on having a reported *E. coli* O157:H7 illness.

For the primary spatial scan statistic model, we used a maximum cluster size of 20% of cases. Statistical significance of the clusters was determined based on Monte Carlo replications under the null. Relative risks presented estimate risk of one's infection being from the given lineage inside the cluster compared with the risk outside that cluster.

We identified 3 statistically significant clusters in which the distribution of cases by phylogenetic lineage varied from the distribution in the rest of the state (Technical Appendix Figure 2). The first cluster ($p = 0.001$) contained 203 cases, was centered in the southwest region of the state, and was characterized by a higher proportion of lineage IIb cases than observed elsewhere in the state (relative risk [RR] 2.59). The second cluster ($p = 0.001$), encompassing the sparsely populated northern reaches of the state, contained 185 cases and had somewhat more Ib (RR 1.37) and rare lineage (RR 1.88) cases and fewer IIb cases (RR 0.29). The final significant cluster ($p = 0.006$) contained 79 cases in the south-central region of the state; lineage IIa was more common than elsewhere in the state (RR 1.70), IIb was uncommon (RR 0.13), and cases due to rare lineages were nearly absent (RR 0). The first cluster, dominated by IIb, and third

cluster, dominated by IIa, recapitulate the results of the kernel estimation maps and, for IIb, the GAM-generated risk surface. The second cluster, dominated by lineage Ib, is larger and centered somewhat further east than the area of segregation identified for Ib by the kernel estimation method, though still similar.

Altering the parameters of the analysis to allow lower or higher percentages of the cases to be included in clusters did not meaningfully affect the position of the clusters identified. We tested allowing clusters up to 50% of cases and 10% of cases. From the former, the main IIb-dominant and Ib/rare-dominant clusters were identified, but the IIa-dominated cluster was not. Limiting clusters to 10% of cases, all 3 clusters identified in the primary analysis were identified but with smaller numbers of included cases.

We detected variant clusters using multinomial spatiotemporal scan statistics, using year as the time scale and allowing up to 50% of the study period in a cluster, as well as purely spatial clusters. We identified 3 statistically significant clusters (Technical Appendix Figure 3). The first ($p = 0.001$) contained 76 cases reported during 2009–2012 in the southwest region of the state and had an elevated risk of lineage IIb (RR 4.45). The second cluster ($p = 0.001$) included 107 cases across the northeast region during 2005–2009. The Ib (RR 1.61) and rare (RR 1.88) lineages were elevated. The third cluster ($p = 0.002$) included only 46 cases reported during 2009–2010, with a predominance of lineage IIb (RR 3.63) and near-absence of IIa (RR 0.09). This cluster included part of Seattle, Washington's largest urban area, and areas immediately south and east.

Secondary Cases

To separate the effect of person-to-person transmission from other potential environmental factors that may result in segregation, we conducted sensitivity analyses after excluding known secondary cases. To be excluded, the most likely source of the infection had to have been identified during the public health investigation as person-to-person, or the notes had to indicate that another individual in the household or childcare situation had previously received such a diagnosis. Based on these criteria, 82 secondary cases were excluded. No meaningful changes in the results were observed. The overall test of spatial segregation was statistically significant using the kernel estimation method ($p = 0.002$) and the nearest-neighbor method ($p < 0.001$). The latitude/longitude smooth of lineage IIb from the multinomial GAM is statistically

significantly different from that of lineage Ib ($p < 0.001$). However, the cluster identified in the southwest region of the state, dominated by lineage IIb, through multinomial spatial scan statistics moved somewhat northward and decreased in size without the secondary cases.

Reporting Bias

We assessed potential reporting bias by county. Reporting of patients who have tested positive is considered near 100% (*11*), but testing intensity may vary by provider. *E. coli* O157:H7 is most often detected by fecal specimen culture, a test that also detects *Campylobacter*, *Salmonella*, and *Shigella*. If providers in an area have heightened awareness of *E. coli* O157:H7 and are more likely to test for it than in other areas, we would expect that detection of these other pathogens would also be higher. There is overlap in the epidemiology of *E. coli* O157:H7, *Campylobacter*, and *Salmonella*, so some correlation is expected. However, risk factors for *Shigella* are generally different (*12*). If there were reporting bias, we would expect this to have the greatest impact on the observed incidence of milder *E. coli* O157:H7 strains.

Case counts by county for 2005–2014 for campylobacteriosis, salmonellosis, and shigellosis were obtained from the Washington State Communicable Disease Reports for 2009 and 2014 (each contained 5 years of data) (*13,14*). We calculated incidence rates using county populations as reported in 2010 U.S. Census TIGER/Line Shapefiles (*15*). Using the GISTools (*16*) package in R, we mapped the incidence quintile of each of the 4 pathogens at the county level for the study period to assess the potential for reporting bias (Technical Appendix Figure 4). Two counties, Yakima and Grant, appear in the uppermost quintile of incidence for each of the 4 diseases. However, incidence of rare lineage *E. coli* O157:H7 in this region is remarkably low (main article Figure 1; Technical Appendix Figure 2). Infections caused by these bacteria are generally milder (main article Table) and would be the type whose numbers would be exaggerated in the presence of heightened testing. Thus, it is unlikely that reporting bias is responsible for the observed results.

Data

Genomic data, with limited metadata, on all isolates used in the study are provided in Technical Appendix 2 (https://wwwnc.cdc.gov/EID/article/23/1/17-0851-Techapp2.xlsx). These include genomic data on all 1,160 *E. coli* O157:H7 isolates from reported, culture-confirmed

cases in Washington state, 2005–2014. Phylogenetic lineage was determined directly using the 48-plex SNP assay developed by Jung et al. (*17*) or was inferred from a typed isolate with the same PFGE profile. Shiga toxin bacteriophage insertion typing and typing for clade according to the method used by Manning et al. (*2*) were conducted on only a subset of isolates. NT, not typed; PFGE, pulsed field gel electrophoresis; SBI, Shiga toxin bacteriophage insertion typing; SDM, Shannon Manning clade/genotype.

References

1. Iyoda S, Manning SD, Seto K, Kimata K, Isobe J, Etoh Y, et al. Phylogenetic clades 6 and 8 of enterohemorrhagic *Escherichia coli* O157:H7 with particular stx subtypes are more frequently found in isolates from hemolytic uremic syndrome patients than from asymptomatic carriers. Open Forum Infect Dis. 2014;1:ofu061. http://dx.doi.org/10.1093/ofid/ofu061

2. Manning SD, Motiwala AS, Springman AC, Qi W, Lacher DW, Ouellette LM, et al. Variation in virulence among clades of *Escherichia coli* O157:H7 associated with disease outbreaks. Proc Natl Acad Sci U S A. 2008;105:4868–73. http://dx.doi.org/10.1073/pnas.0710834105

3. Pianciola L, Chinen I, Mazzeo M, Miliwebsky E, González G, Müller C, et al. Genotypic characterization of *Escherichia coli* O157:H7 strains that cause diarrhea and hemolytic uremic syndrome in Neuquén, Argentina. Int J Med Microbiol. 2014;304:499–504. http://dx.doi.org/10.1016/j.ijmm.2014.02.011

4. Diggle PJ, Zheng P, Durr P. Nonparametric estimation of spatial segregation in a multivariate point process: bovine tuberculosis in Cornwall, UK. Appl Stat. 2005;54:645–58. http://dx.doi.org/10.1111/j.1467-9876.2005.05373.x

5. R Core Team. R: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2015.

6. Zheng P, Diggle PJ. Spatialkernel: nonparametric estimation of spatial segregation in a multivariate point process; R package version 0.4-19. 2013 [cited 2015 May 27]. https://CRAN.R-project.org/package=spatialkernel

7. Dixon PM. Nearest-neighbor contingency table analysis of spatial segregation for several species. Écoscience. 2002;9:142–51. http://dx.doi.org/10.1080/11956860.2002.11682700

8. de la Cruz M. Metodos Para Analizar Datos Puntuales. In: Maestre, FT, Escudero, A, Bonet, A, editors. Introduccion Al Analisis Espacial De Datos En Ecologia Y Ciencias Ambientales: Metodos Y Aplicaciones. Madrid: Asociacion Espanola de Ecologia Terrestre, Universidad Rey Juan Carlos and Caja de Ahorros del Mediterraneo; 2008. p. 76–127.

9. Jung I, Kulldorff M, Richard OJ. A spatial scan statistic for multinomial data. Stat Med. 2010;29:1910–8. http://dx.doi.org/10.1002/sim.3951

10. Kulldorff M, Information Management Services I. Satscan^{tm} V8.0: Software for the spatial and space-time scan statistics. 2009 [cited 2015 May 27]. http://www.satscan.org/

11. Scallan E, Hoekstra RM, Angulo FJ, Tauxe RV, Widdowson MA, Roy SL, et al. Foodborne illness acquired in the United States—major pathogens. Emerg Infect Dis. 2011;17:7–15. http://dx.doi.org/10.3201/eid1701.P11101

12. Denno DM, Keene WE, Hutter CM, Koepsell JK, Patnode M, Flodin-Hursh D, et al. Tri-county comprehensive assessment of risk factors for sporadic reportable bacterial enteric infection in children. J Infect Dis. 2009;199:467–76. http://dx.doi.org/10.1086/596555

13. Communicable Disease Epidemiology Section. Communicable Disease Report 2014. Shoreline, WA: Washington State Department of Health; 2014. p. 95.

14. Communicable Disease Epidemiology Section. Communicable Disease Report 2009. Shoreline, WA: Washington State Department of Health; 2009. p. 91.

15. United States Census Bureau. 2012 Tiger/Line Shapefiles (Machine-Readable Data Files). Washington, DC: The Bureau; 2012.

16. Brundsdon C, Chen H. GIStools: some further GIS capabilities for R; R package version 0.7-4. 2014 [cited 2015 May 27]. https://CRAN.R-project.org/package=GISTools

17. Jung WK, Bono JL, Clawson ML, Leopold SR, Shringi S, Besser TE. Lineage and genogroup-defining single nucleotide polymorphisms of *Escherichia coli* O157:H7. Appl Environ Microbiol. 2013;79:7036–41. http://dx.doi.org/10.1128/AEM.02173-13

Technical Appendix Table 1. Association of known risk factors with phylogenetic lineage*

Variable	Statewide frequency	Statewide OR (95% CI)	Southwest region (n = 234) OR (95% CI)	Northwest region (n = 289) OR (95% CI)	South-central region (n = 109) OR (95% CI)
Hispanic ethnicity (vs. non-Hispanic)					
Lineage Ib	46/372	Ref	Ref	Ref	Ref
Lineage IIa	32/197	1.13 (0.67, 1.91)	0.3 (0.03, 2.86)	2.79 (0.66, 11.83)	0.87 (0.33, 2.25)
Lineage IIb	19/152	1.13 (0.61, 2.11)	0.99 (0.3, 3.33)	3.24 (0.62, 16.86)	0.73 (0.12, 4.37)
Rare lineage	6/42	1.21 (0.46, 3.15)	8.15 (0.89, 75.06)	1.98 (0.18, 21.31)	0 (0, Inf)†
American Indian (vs. white race)‡					
Lineage Ib	5/377	Ref	Ref	Ref	Ref
Lineage IIa	7/196	**3.82 (1.13, 12.95)§**	NA	NA	NA
Lineage IIb	0/148	0 (0, Inf)†	NA	NA	NA
Rare lineage	0/40	0 (0, Inf)†	NA	NA	NA
Asian race (vs. white race)‡					
Lineage Ib	24/377	Ref	Ref	Ref	Ref
Lineage IIa	7/196	0.53 (0.22, 1.28)	NA	NA	NA
Lineage IIb	19/148	**2.03 (1.02, 4.01)§**	NA	NA	NA
Rare lineage	2/40	0.72 (0.16, 3.22)	NA	NA	NA
Black race (vs. white race)‡					
Lineage Ib	12/377	Ref	Ref	Ref	Ref
Lineage IIa	5/196	0.81 (0.27, 2.43)	NA	NA	NA
Lineage IIb	5/148	1.02 (0.34, 3.06)	NA	NA	NA
Rare lineage	0/40	0 (0, Inf)†	NA	NA	NA
Other/multiple race (vs. white race)‡					
Lineage Ib	16/377	Ref	Ref	Ref	Ref
Lineage IIa	9/196	0.94 (0.39, 2.23)	NA	NA	NA
Lineage IIb	11/148	1.59 (0.69, 3.68)	NA	NA	NA
Rare lineage	1/40	0.55 (0.07, 4.32)	NA	NA	NA
Contact with a laboratory-confirmed case					
Lineage Ib	59/531	Ref	Ref	Ref	Ref
Lineage IIa	39/228	1.34 (0.84, 2.15)	0.88 (0.3, 2.6)	1.48 (0.63, 3.49)	0.99 (0.25, 3.96)
Lineage IIb	43/176	**1.96 (1.21, 3.16)¶**	**2.7 (1.15, 6.31)§**	2.03 (0.78, 5.25)	2.74 (0.44, 17.21)
Rare lineage	3/60	0.41 (0.12, 1.37)	0.42 (0.05, 3.82)	0.39 (0.05, 3.24)	0 (0, Inf)†
Epidemiologic link to a confirmed or probable case					
Lineage Ib	74/522	Ref	Ref	Ref	Ref
Lineage IIa	41/221	1.25 (0.80, 1.96)	1.07 (0.37, 3.05)	0.97 (0.42, 2.25)	0.99 (0.24, 3.98)
Lineage IIb	51/172	**1.94 (1.24, 3.03)¶**	2.17 (0.94, 4.98)	1.41 (0.56, 3.55)	4.72 (0.85, 26.07)
Rare lineage	3/60	0.32 (0.10, 1.06)	0.33 (0.04, 2.95)	0.29 (0.04, 2.39)	0 (0, Inf)†
Underlying illness					
Lineage Ib	66/530	Ref	Ref	Ref	Ref
Lineage IIa	27/233	1.20 (0.70, 2.06)	2.87 (0.86, 9.61)	0.83 (0.2, 3.37)	4.07 (0.5, 33.02)
Lineage IIb	19/184	1.11 (0.61, 2.01)	1.17 (0.36, 3.77)	0.73 (0.15, 3.59)	6.07 (0.33, 111.66)
Rare lineage	2/62	**0.19 (0.04, 0.85)§**	0.59 (0.06, 5.84)	0.42 (0.05, 3.73)	0 (0, Inf)†
Contact with diapered or incontinent child or adult					
Lineage Ib	122/545	Ref	Ref	Ref	Ref
Lineage IIa	65/231	1.10 (0.75, 1.61)	0.91 (0.37, 2.22)	0.94 (0.42, 2.1)	1.43 (0.54, 3.79)
Lineage IIb	60/187	1.28 (0.86, 1.91)	1.57 (0.76, 3.26)	1.58 (0.67, 3.73)	0.82 (0.13, 5.17)
Rare lineage	8/62	0.53 (0.24, 1.16)	1.44 (0.36, 5.72)	0.19 (0.02, 1.52)	0.69 (0.06, 7.7)
Attends childcare or preschool					
Lineage Ib	39/523	Ref	Ref	Ref	Ref
Lineage IIa	22/235	DNC	2.7 (0.68, 10.64)	1.7 (0.42, 6.86)	1.19 (0.21, 6.56)
Lineage IIb	27/181	DNC	**3.17 (1.03, 9.7)§**	2.16 (0.55, 8.57)	0 (0, Inf)†
Rare lineage	0/59	DNC	0 (0, Inf)†	0 (0, Inf)†	0 (0, Inf)†
Employed as a healthcare worker					
Lineage Ib	17/525	Ref	Ref	Ref	Ref
Lineage IIa	8/232	DNC	3.06 (0.44, 21.55)	0.7 (0.06, 8.42)	0 (0, Inf)†
Lineage IIb	7/182	DNC	0.71 (0.06, 8.23)	1.52 (0.15, 15.38)	2.41 (0.18, 33.1)
Rare lineage	1/62	DNC	0 (0, Inf)†	0 (0, Inf)†	0 (0, Inf)†
Employed as a food worker					
Lineage Ib	18/539	Ref	Ref	Ref	Ref
Lineage IIa	12/244	1.64 (0.74, 3.59)	1.4 (0.1, 19.6)	1.58 (0.45, 5.56)	0 (0, Inf)†
Lineage IIb	4/188	0.74 (0.24, 2.28)	1.61 (0.21, 12.62)	0.53 (0.06, 4.44)	0 (0, Inf)†
Rare lineage	2/60	0.99 (0.22, 4.41)	0 (0, Inf)†	1.11 (0.13, 9.77)	0 (0, Inf)†

Variable	Statewide frequency	Statewide OR (95% CI)	Southwest region (n = 234) OR (95% CI)	Northwest region (n = 289) OR (95% CI)	South-central region (n = 109) OR (95% CI)
Works with animals or animal products					
Lineage Ib	24/524	Ref	Ref	Ref	Ref
Lineage IIa	5/196	0.46 (0.16, 1.27)	0 (0, Inf)†	0.31 (0.04, 2.57)	0.87 (0.12, 6.08)
Lineage IIb	5/163	0.84 (0.30, 2.40)	0.77 (0.11, 5.45)	0 (0, Inf)†	2.17 (0.19, 24.56)
Rare lineage	3/53	1.14 (0.33, 4.00)	2.87 (0.25, 33.45)	1.73 (0.32, 9.22)	0 (0, Inf)†
Any contact with animals					
Lineage Ib	300/521	Ref	Ref	Ref	Ref
Lineage IIa	115/200	0.81 (0.57, 1.15)	0.84 (0.34, 2.09)	0.56 (0.26, 1.24)	0.48 (0.18, 1.3)
Lineage IIb	90/167	0.78 (0.54, 1.14)	1.9 (0.89, 4.06)	0.48 (0.2, 1.15)	**0.16 (0.03, 0.88)§**
Rare lineage	27/52	0.8 (0.44, 1.45)	Inf (0, Inf)†	0.73 (0.24, 2.26)	0.3 (0.02, 3.63)
Contact with cattle, cows, or calves					
Lineage Ib	63/471	Ref	Ref	Ref	Ref
Lineage IIa	30/188	1.06 (0.64, 1.78)	1.11 (0.32, 3.81)	0.68 (0.26, 1.78)	1.06 (0.36, 3.12)
Lineage IIb	13/151	0.59 (0.3, 1.14)	1.04 (0.38, 2.84)	0.14 (0.02, 1.07)	0 (0, Inf)†
Rare lineage	7/49	1.19 (0.5, 2.81)	0.95 (0.1, 8.81)	0.92 (0.24, 3.54)	0 (0, Inf)†
Case or household member lives or works on a farm or dairy					
Lineage Ib	67/526	Ref	Ref	Ref	Ref
Lineage IIa	24/191	0.86 (0.50, 1.46)	0.47 (0.09, 2.5)	0.96 (0.37, 2.44)	1.06 (0.36, 3.13)
Lineage IIb	15/169	0.67 (0.35, 1.27)	1.62 (0.58, 4.48)	0 (0, Inf)†	0.33 (0.04, 2.95)
Rare lineage	7/53	1.08 (0.47, 2.52)	1.35 (0.14, 13)	0.99 (0.26, 3.82)	1.39 (0.11, 17.56)
Visited a zoo, farm, fair, or pet shop					
Lineage Ib	99/526	Ref	Ref	Ref	Ref
Lineage IIa	49/200	1.31 (0.86, 2)	1.59 (0.61, 4.17)	0.93 (0.41, 2.12)	1 (0.28, 3.53)
Lineage IIb	25/166	**0.59 (0.35, 1)§**	0.88 (0.4, 1.94)	**0.17 (0.04, 0.78)§**	0 (0, Inf)†
Rare lineage	11/53	1.11 (0.53, 2.33)	0.52 (0.06, 4.65)	0.6 (0.16, 2.32)	2.65 (0.2, 34.74)
Recreational water exposure					
Lineage Ib	130/548	Ref	Ref	Ref	Ref
Lineage IIa	57/229	0.96 (0.65, 1.41)	0.51 (0.18, 1.45)	0.53 (0.24, 1.17)	0.79 (0.25, 2.56)
Lineage IIb	38/174	0.82 (0.53, 1.27)	**0.38 (0.16, 0.93)§**	0.44 (0.16, 1.24)	**6.39 (1.09, 37.47)§**
Rare lineage	12/60	0.79 (0.40, 1.57)	0.66 (0.13, 3.41)	0.77 (0.22, 2.73)	1.41 (0.12, 16.12)
Drank untreated/unchlorinated water					
Lineage Ib	61/531	Ref	Ref	Ref	Ref
Lineage IIa	29/219	0.96 (0.58, 1.57)	**4.49 (1.48, 13.57)¶**	0.89 (0.27, 2.87)	**0.16 (0.04, 0.63)¶**
Lineage IIb	26/169	1.27 (0.74, 2.16)	**3.76 (1.38, 10.28)¶**	1.5 (0.44, 5.15)	0.27 (0.03, 2.38)
Rare lineage	7/53	1.14 (0.49, 2.66)	1.68 (0.29, 9.69)	2.14 (0.41, 11.07)	0 (0, Inf)†
Well is source of drinking water					
Lineage Ib	136/559	Ref	Ref	Ref	Ref
Lineage IIa	59/236	0.91 (0.62, 1.32)	1.1 (0.47, 2.54)	1.1 (0.5, 2.41)	0.47 (0.19, 1.17)
Lineage IIb	35/186	0.77 (0.48, 1.21)	1.06 (0.52, 2.12)	0.7 (0.24, 2)	**0.08 (0.01, 0.72)§**
Rare lineage	14/62	0.87 (0.46, 1.65)	0.49 (0.11, 2.09)	1.13 (0.33, 3.84)	0.18 (0.02, 1.73)
Consumed food from a restaurant					
Lineage Ib	384/505	Ref	Ref	Ref	Ref
Lineage IIa	166/216	1.22 (0.81, 1.83)	1.82 (0.69, 4.81)	0.93 (0.4, 2.17)	0.66 (0.25, 1.72)
Lineage IIb	132/171	1.09 (0.7, 1.68)	1.09 (0.5, 2.39)	0.72 (0.29, 1.78)	Inf (0, Inf)†
Rare lineage	43/54	1.23 (0.61, 2.49)	0.74 (0.19, 2.82)	0.82 (0.24, 2.79)	1.61 (0.15, 17.53)
Consumed food from a group meal					
Lineage Ib	144/531	Ref	Ref	Ref	Ref
Lineage IIa	65/227	1.1 (0.77, 1.59)	0.53 (0.19, 1.48)	1.56 (0.72, 3.39)	0.73 (0.28, 1.92)
Lineage IIb	59/179	1.24 (0.84, 1.82)	1.18 (0.58, 2.4)	**2.45 (1.06, 5.71)§**	0.27 (0.03, 2.52)
Rare lineage	17/58	1.16 (0.64, 2.13)	0.58 (0.12, 2.86)	**3.1 (1.02, 9.4)§**	0.46 (0.04, 4.8)
Handled raw meat					
Lineage Ib	122/542	Ref	Ref	Ref	Ref
Lineage IIa	43/226	0.86 (0.54, 1.38)	1.21 (0.4, 3.64)	0.75 (0.3, 1.88)	1.5 (0.37, 6.14)
Lineage IIb	31/182	0.92 (0.55, 1.53)	1.41 (0.55, 3.61)	0.23 (0.05, 1.08)	0.51 (0.07, 3.9)
Rare lineage	15/62	1.09 (0.54, 2.17)	1.47 (0.33, 6.47)	0.62 (0.15, 2.49)	2.14 (0.17, 27.6)
Consumed meat					
Lineage Ib	314/521	Ref	Ref	Ref	Ref
Lineage IIa	138/223	1.09 (0.77, 1.53)	1.09 (0.48, 2.47)	1.33 (0.59, 2.99)	1.45 (0.58, 3.62)
Lineage IIb	106/175	1.07 (0.74, 1.55)	1.25 (0.64, 2.44)	1.83 (0.63, 5.37)	1.3 (0.28, 6.08)
Rare lineage	31/56	0.75 (0.43, 1.33)	0.59 (0.16, 2.13)	0.46 (0.15, 1.42)	0.77 (0.11, 5.23)
Consumed ground beef					
Lineage Ib	331/539	Ref	Ref	Ref	Ref
Lineage IIa	132/229	0.85 (0.61, 1.18)	0.94 (0.39, 2.3)	0.88 (0.43, 1.8)	0.82 (0.32, 2.09)
Lineage IIb	103/180	0.85 (0.59, 1.22)	0.87 (0.43, 1.76)	**0.27 (0.11, 0.65)¶**	0.82 (0.18, 3.78)
Rare lineage	31/57	0.76 (0.44, 1.34)	1.52 (0.29, 8.01)	0.6 (0.22, 1.69)	0.31 (0.04, 2.14)
Consumed intact beef					
Lineage Ib	283/462	Ref	Ref	Ref	Ref

Variable	Statewide frequency	Statewide OR (95% CI)	Southwest region (n = 234) OR (95% CI)	Northwest region (n = 289) OR (95% CI)	South-central region (n = 109) OR (95% CI)
Lineage IIa	116/185	1.07 (0.74, 1.56)	0.55 (0.2, 1.5)	0.87 (0.4, 1.89)	2.81 (0.85, 9.3)
Lineage IIb	90/156	0.86 (0.58, 1.28)	0.96 (0.42, 2.17)	**0.35 (0.14, 0.87)§**	1.36 (0.22, 8.52)
Rare lineage	29/46	1.17 (0.61, 2.27)	2.77 (0.3, 25.43)	1.54 (0.43, 5.51)	0.31 (0.01, 7.79)
Consumed venison or other wild game meat					
Lineage Ib	15/521	Ref	Ref	Ref	Ref
Lineage IIa	3/195	0.37 (0.08, 1.68)	0 (0, Inf)†	0 (0, Inf)†	0 (0, Inf)†
Lineage IIb	10/169	1.97 (0.81, 4.79)	1.35 (0.4, 4.58)	1.22 (0.13, 11.12)	0 (0, Inf)†
Rare lineage	5/53	**3.56 (1.23, 10.32)§**	1.56 (0.16, 14.98)	3.56 (0.58, 21.96)	**34.96 (1.03, 1187.37)§**
Consumed raw milk					
Lineage Ib	16/551	Ref	Ref	Ref	Ref
Lineage IIa	6/232	0.82 (0.3, 2.23)	4.04 (0.22, 75.92)	0.38 (0.04, 3.72)	0 (0, Inf)†
Lineage IIb	18/183	**2.46 (1.15, 5.28)§**	**17.33 (2.05, 146.5)¶**	0 (0, Inf)†	24.32 (0.81, 726.95)
Rare lineage	1/60	0.63 (0.08, 4.88)	0 (0, Inf)†	0 (0, Inf)†	0 (0, Inf)†
Consumed unpasteurized juice					
Lineage Ib	11/496	Ref	Ref	Ref	Ref
Lineage IIa	3/219	0.34 (0.09, 1.27)	0.8 (0.11, 6.04)	0 (0, Inf)†	0 (0, Inf)†
Lineage IIb	7/163	1.53 (0.55, 4.29)	0.6 (0.09, 4.03)	7.08 (0.37, 137.1)	5.9 (0.35, 100.4)
Rare lineage	3/55	2.31 (0.61, 8.78)	2.39 (0.21, 27.47)	**23.08 (1.52, 351.69)§**	0 (0, Inf)†
Consumed raw fruits or vegetables					
Lineage Ib	435/514	Ref	Ref	Ref	Ref
Lineage IIa	184/205	**1.81 (1.05, 3.11)§**	6.88 (0.84, 56.67)	2.55 (0.52, 12.41)	1.34 (0.43, 4.16)
Lineage IIb	144/170	1.25 (0.74, 2.1)	1.51 (0.62, 3.64)	0.78 (0.23, 2.6)	1.97 (0.2, 19.15)
Rare lineage	43/48	1.5 (0.57, 4)	Inf (0, Inf)†	2.11 (0.25, 17.82)	0.37 (0.02, 5.85)
Consumed sprouts					
Lineage Ib	22/537	Ref	Ref	Ref	Ref
Lineage IIa	12/231	1.45 (0.68, 3.11)	1.87 (0.23, 15.21)	2.98 (0.57, 15.62)	Inf (0, Inf)†
Lineage IIb	12/180	2 (0.94, 4.27)	1.11 (0.17, 7.45)	**5.17 (1.04, 25.74)§**	0.5 (0, Inf)
Rare lineage	4/57	1.94 (0.64, 5.94)	0 (0, Inf)†	**7.32 (1.11, 48.28)§**	0.24 (0, Inf)
Consumed fresh herbs					
Lineage Ib	102/524	Ref	Ref	Ref	Ref
Lineage IIa	44/216	0.83 (0.54, 1.27)	0.95 (0.32, 2.79)	0.88 (0.37, 2.1)	**0.19 (0.04, 0.77)§**
Lineage IIb	35/178	1.01 (0.64, 1.6)	0.78 (0.29, 2.13)	1.51 (0.59, 3.85)	0.39 (0.04, 3.57)
Rare lineage	9/56	0.7 (0.32, 1.55)	0 (0, Inf)†	1.11 (0.29, 4.3)	0.39 (0.03, 4.47)
Traveled outside the state, the country, or usual routine					
Lineage Ib	143/571	Ref	Ref	Ref	Ref
Lineage IIa	52/246	0.78 (0.53, 1.13)	0.45 (0.17, 1.19)	0.37 (0.14, 1)	1.09 (0.34, 3.54)
Lineage IIb	54/197	1.08 (0.74, 1.59)	0.86 (0.44, 1.7)	1.71 (0.73, 4)	1.53 (0.26, 9.01)
Rare lineage	26/64	**2.03 (1.17, 3.50)§**	0.66 (0.16, 2.65)	**3.72 (1.27, 10.87)§**	**7.45 (1.03, 54.07)§**

*All analyses are multinomial logistic regression, using lineage Ib as the reference group, adjusted for age, sex, and year. The statewide analysis was conducted using a generalized additive model to additionally adjust for latitude and longitude using a thin plate spline bivariate smooth. Statistically significant results are shown in bold text. “Rare lineage” includes 12 different clinically rare lineages. CI, confidence interval; DNC, did not converge; Inf, infinity; NA, not applicable; OR, odds ratio; Ref, reference

†Odds ratios of 0 are reported where 0 cases of the lineage under analysis existed in the category. Odds ratios of infinity are reported where 0 cases of the reference lineage (Ib) existed in the category. Confidence intervals were not estimated for these ORs, indicated by (0, Inf).

‡ Analyses marked NA could not be performed or were considered unreliable because of sparse data in these categories. Not all models converged because of sparse data in some categories.

§ $p < 0.05$

¶ $p < 0.01$

Technical Appendix Table 2. Multinomial generalized additive model sensitivity analysis

Model	Latitude/longitude p-value	AIC
Bivariate thin plate regression spline model for latitude/longitude, age, and sex covariates*	IIa: 0.127	1337
	IIb: <0.001	
	Rare: 0.692	
Intercept only	NA	1396
Univariate thin plate regression spline models for latitude and longitude	IIa latitude: 0.022	1338
	IIa longitude: 0.967	
	IIb latitude: <0.001	
	IIb longitude: <0.001	
	Rare latitude: 0.399	
	Rare longitude: 0.734	
Bivariate thin plate regression spline model for latitude/longitude	IIa: 0.071	1340
	IIb: <0.001	
	Rare: 0.688	
Bivariate thin plate regression spline model for latitude/longitude, age and sex covariates, basis dimension doubled	IIa: 0.127	1336
	IIb: <0.001	
	Rare: 0.691	
Cubic regression spline models for latitude and longitude, age and sex covariates	IIa latitude: 0.042	1336
	IIa longitude: 0.845	
	IIb latitude: <0.001	
	IIb longitude: <0.001	
	Rare latitude: 0.425	
	Rare longitude: 0.646	
Bivariate tensor product spline model for latitude/longitude, age and sex covariates	IIa: 0.077	1338
	IIb: <0.001	
	Rare: 0.860	
Bivariate thin plate regression spline model for latitude/longitude, age and sex covariates, using lineage IIa as the comparator instead of Ib	Ib: 0.127	1969
	IIb: <0.001	
	Rare: 0.189	

Model	Latitude/longitude p-value	AIC
Bivariate thin plate regression spline model for latitude/longitude; age, sex, and year covariates	IIa: 0.104 IIb: <0.001 Rare: 0.739	1273
Thin plate regression spline models for latitude/longitude (bivariate) and year (univariate), age and sex covariates	IIa: 0.116 IIb: <0.001 Rare: 0.730	1237
Trivariate thin plate regression spline model for latitude/longitude/year, age and sex covariates	IIa latitude/longitude/year: <0.001 IIb latitude/longitude/year: <0.001 Rare latitude/longitude/year: 0.475	1174

*Primary model. AIC, Akaike information criterion; NA, not applicable

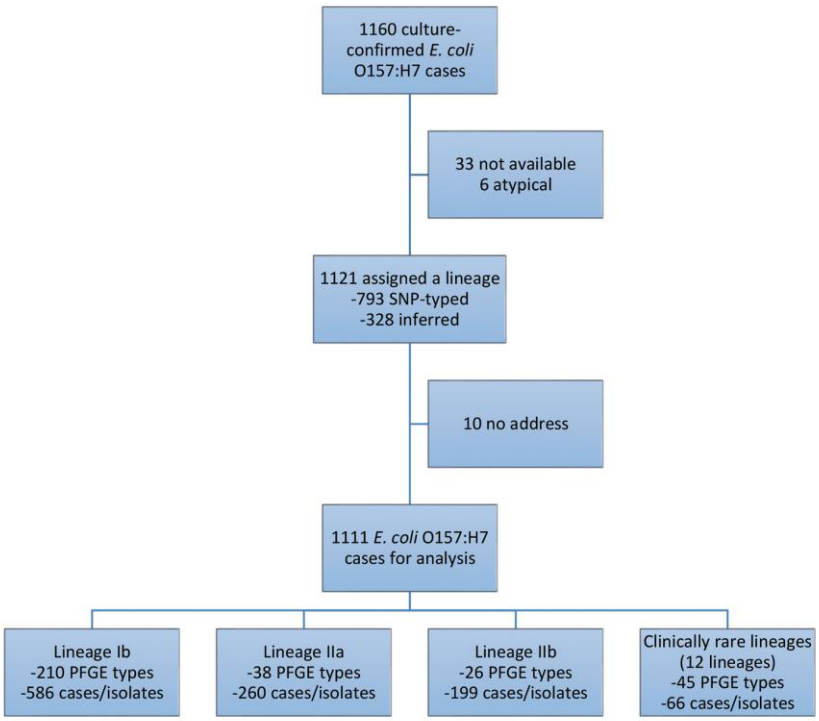
Technical Appendix Table 3. Dixon nearest-neighbor contingency table analysis of spatial segregation

Lineage	df*	χ^2	p-value
Overall	12	96.19	<0.001
Ib	3	8.02	0.046
IIa	3	15.08	0.002
IIb	3	75.61	<0.001
Rare	3	4.04	0.257

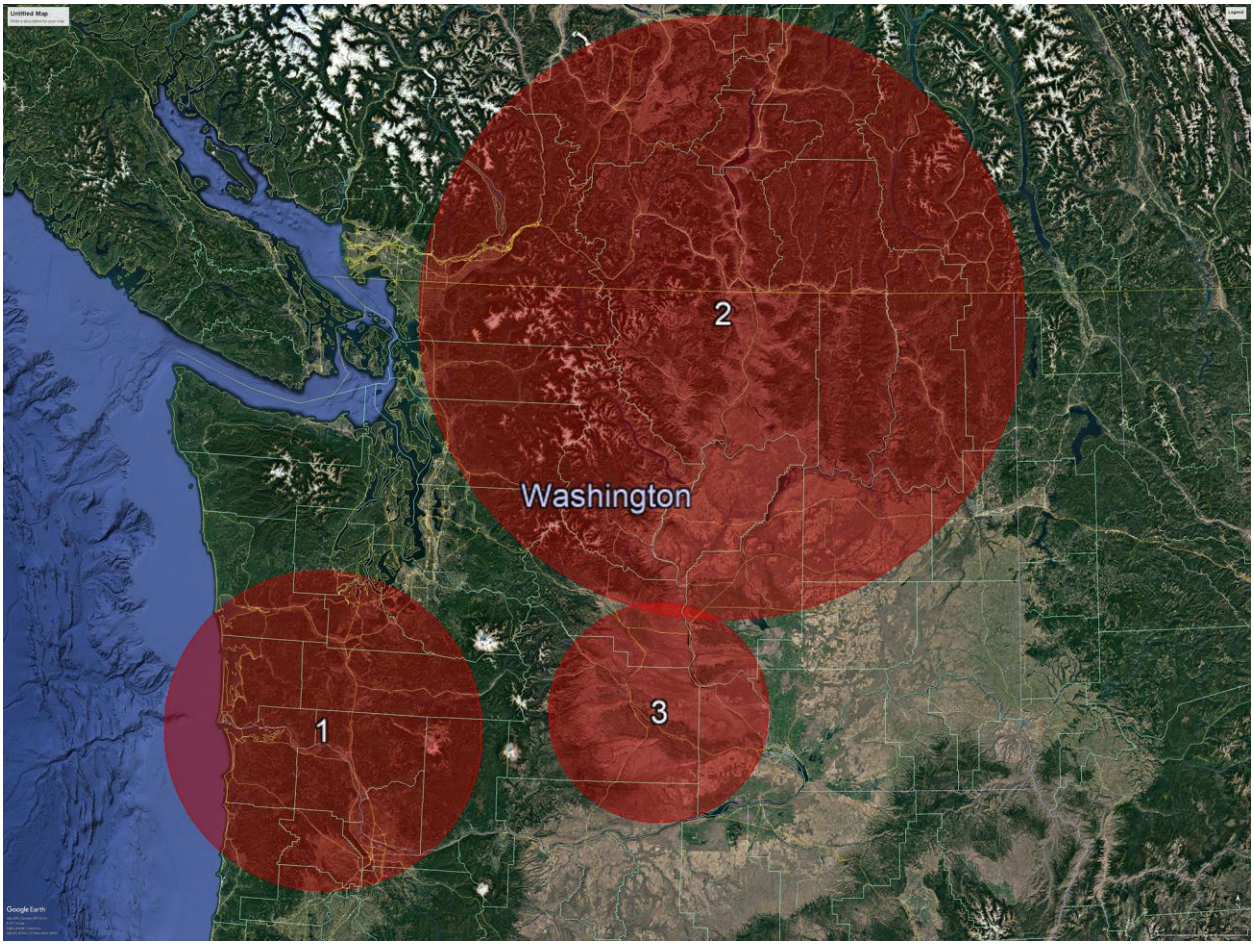
* df, degrees of freedom

Technical Appendix Table 4. Pairwise segregation of lineages using Dixon's nearest-neighbor contingency table method

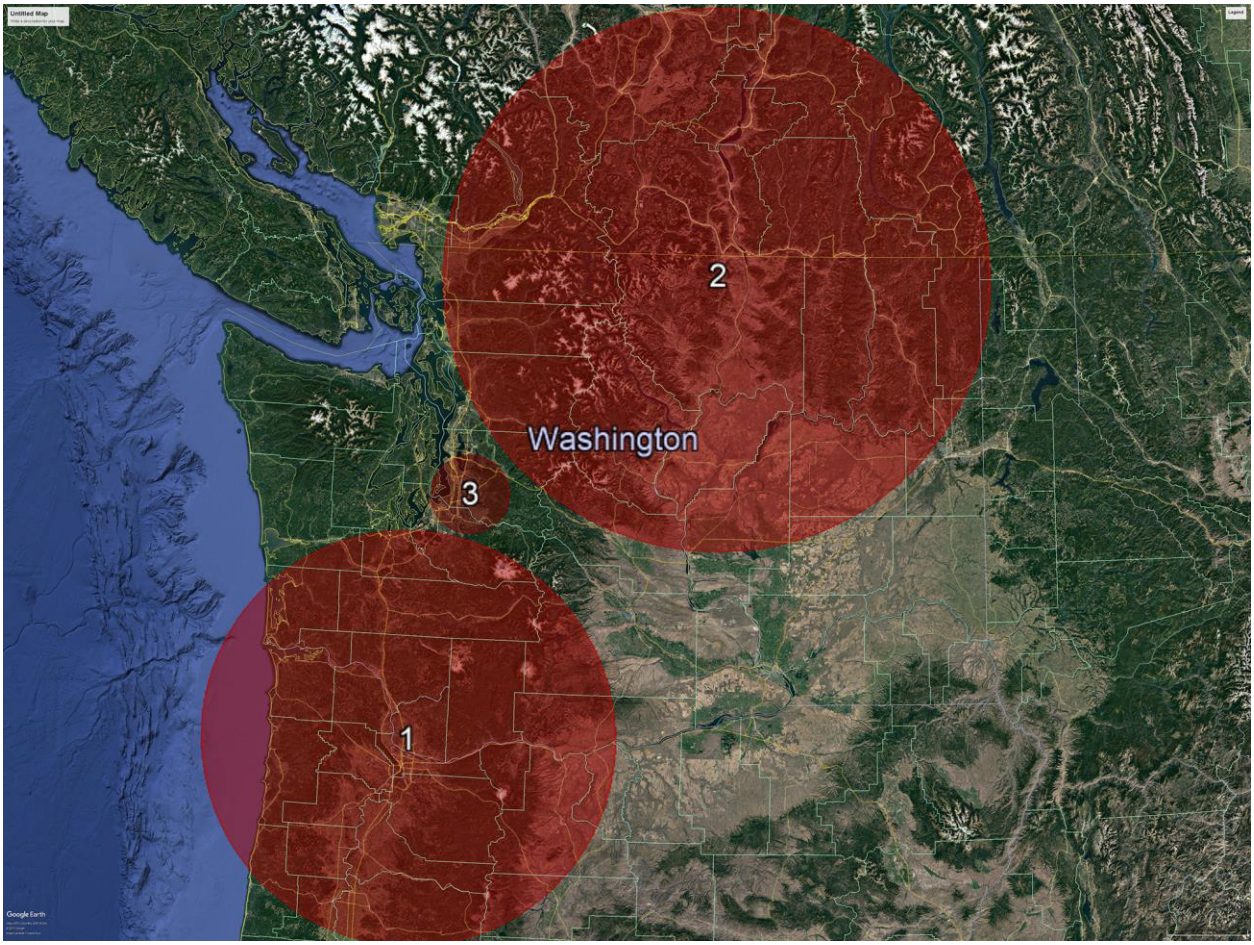
From	To	Observed Count	Expected Count	*S*	*Z*-score	p-value
Ib	Ib	343	308.84	0.10	2.61	0.009
Ib	IIa	115	137.26	−0.10	−2.19	0.028
Ib	IIb	92	105.06	−0.07	−1.44	0.150
Ib	Rare	36	34.84	0.02	0.21	0.832
IIa	Ib	122	137.26	−0.10	−1.80	0.072
IIa	IIa	90	60.67	0.24	3.61	<0.001
IIa	IIb	40	46.61	−0.08	−1.08	0.280
IIa	Rare	8	15.46	−0.30	−2.00	0.046
IIb	Ib	80	105.06	−0.22	−3.42	<0.001
IIb	IIa	24	46.61	−0.35	−3.80	<0.001
IIb	IIb	91	35.50	0.59	8.50	<0.001
IIb	Rare	4	11.83	−0.49	−2.39	0.017
Rare	Ib	43	34.84	0.22	1.98	0.047
Rare	IIa	11	15.46	−0.18	−1.30	0.195
Rare	IIb	9	11.83	−0.14	−0.91	0.362
Rare	Rare	3	3.86	−0.12	−0.36	0.717



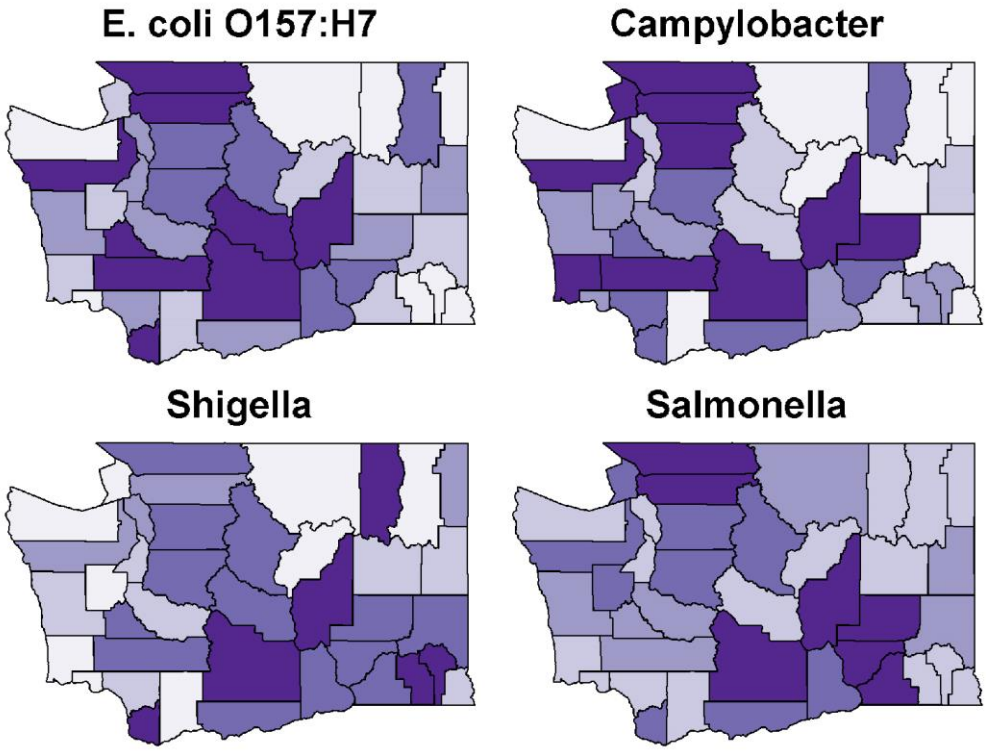
Technical Appendix Figure 1. Of the 1,160 culture-confirmed *E. coli* O157:H7 cases reported in Washington state during 2005–2014, 1,111 were included in the analysis. Isolates from these 1,111 cases spanned 15 phylogenetic lineages using the 48-plex single nucleotide polymorphism assay developed by Jung et al. (*17*). Three lineages, Ib, IIa, and IIb, constituted 94% of isolates. Isolates from the remaining 12 lineages were grouped into a "clinically rare" group. XbaI pulsed field gel electrophoresis (PFGE) types were determined, and all isolates of a given PFGE type belonged to the same phylogenetic lineage. The number of PFGE types and case isolates belonging to each lineage are shown.



Technical Appendix Figure 2. Statistically significant clusters of variant phylogenetic lineage. Multinomial spatial scan statistics were used to identify clusters in which the distribution of lineages varied from that of the rest of the state. Clusters were restricted to a maximum of 20% of cases. Cluster 1: 203 cases; Ib relative risk (RR) = 0.66, IIa RR = 0.94, IIb RR = 2.59, Rare RR = 0.80; p = 0.001. Cluster 2: 185 cases; Ib RR = 1.37, IIa RR = 0.65, IIb RR = 0.29, Rare RR = 1.88; p = 0.001. Cluster 3: 79 cases; Ib RR = 1.14, IIa RR = 1.70, IIb RR = 0.13, Rare RR = 0; p = 0.006.



Technical Appendix Figure 3. Statistically significant space-time clusters of variant phylogenetic lineage. Multinomial spatiotemporal scan statistics were used to identify clusters in which the distribution of lineages varied from that of the rest of the state during years outside the cluster. Clusters were restricted to a maximum of 20% of cases and 50% of the study window. Cluster 1: 2009–2012; 76 cases; Ib relative risk (RR) = 0.28, IIa RR = 0.49, IIb RR = 4.45, Rare RR = 1.36; p = 0.001. Cluster 2: 2005–2009; 107 cases; Ib RR = 1.61, IIa RR = 0.22, IIb RR = 0.19, Rare RR = 1.88; p = 0.001. Cluster 3: 2009–2010; 46 cases; Ib RR = 0.65, IIa RR = 0.09, IIb RR = 3.63, Rare RR = 0.72; p = 0.002.



Technical Appendix Figure 4. Incidence rate quintiles by county of reported *E. coli* O157, *Campylobacter*, *Shigella,* and *Salmonella*, 2005–2014. Tests are routinely performed for these 4 pathogens simultaneously, and uniformly high rates may suggest higher testing intensity in a county.