Persistent cross-species transmission systems dominate Shiga toxin-producing *Escherichia coli* O157:H7 epidemiology in a high incidence region: a genomic epidemiology study

Gillian A.M. Tarr, PhD,[1] Linda Chui, PhD,[2,3] Kim Stanford, PhD,[4] Emmanuel W. Bumunang, PhD,[4] Rahat Zaheer, PhD,[5] Vincent Li, MSc,[2] Stephen B. Freedman, MDCM,[6] Chad R. Laing, PhD,[7] Tim A. McAllister, PhD[5]

[1] Division of Environmental Health Sciences, School of Public Health, University of Minnesota, MMC807, Room 1240 Mayo, 420 Delaware Street SE, Minneapolis, MN, 55455, USA

[2] Alberta Precision Laboratories, Alberta Public Health, Room 1B2.19 Walter Mackenzie Health Sciences Centre, 8440-112th Street, Edmonton, Alberta, T6G 2J2, Canada

[3] Department of Laboratory Medicine and Pathology, University of Alberta, Edmonton, Alberta, Canada

[4] Department of Biological Sciences, University of Lethbridge, 4401 University Drive, Lethbridge, Alberta, T1K 3M4, Canada

[5] Agriculture and Agri-Food Canada, Lethbridge Research and Development Centre, 5403 – 1 Avenue South. P.O. Box 3000, Lethbridge, Alberta, T1J 4B1, Canada

[6] Sections of Pediatric Emergency Medicine and Gastroenterology, Department of Pediatrics, Alberta Children's Hospital and Alberta Children's Hospital Research Institute, Cumming School of Medicine, University of Calgary, C4-634 28 Oki Drive NW, Calgary, Alberta, T3B 6A8, Canada

[7] National Center for Animal Diseases Lethbridge Laboratory, Canadian Food Inspection Agency, 225090 Township Road 91, Lethbridge County, Alberta, T1J 5R7, Canada

Corresponding author: Gillian A.M. Tarr; Division of Environmental Health Sciences, School of Public Health, University of Minnesota, MMC807, Room 1240 Mayo, 420 Delaware Street SE, Minneapolis, MN, 55455, USA; gtarr@umn.edu; 612-626-9308

Keywords: *E. coli* O157, STEC, genomic epidemiology, zoonosis

ABSTRACT

BACKGROUND

Several areas of the world suffer notably high incidence of Shiga toxin-producing *Escherichia coli*, among them Alberta, Canada. We assessed the role of persistent cross-species transmission systems in Alberta's *E. coli* O157:H7 epidemiology.

METHODS

We sequenced and assembled 229 *E. coli* O157:H7 isolates originating from collocated cattle (n=108) and human (n=121) populations from 2007-2015 in Alberta. We constructed a timed phylogeny using BEAST2 using a structured coalescent model. We then extended the tree with human isolates through 2019 (n=432) to assess the long-term disease impact of local persistent lineages. Shiga toxin gene (*stx*) profile was determined for all isolates.

RESULTS

During 2007 to 2015, we estimated 107 (95% HPD 101, 111) human lineages arose from cattle lineages, and 31 (95% HPD 22, 43) from other human lineages; i.e., 77.5% of human lineages arose from cattle lineages. We identified 11 persistent lineages local to Alberta, which were associated with 36.4% (95% CI 27.8%, 45.6%) of human isolates. Of 115 isolates in local persistent lineages, 6.1% carried only *stx2a* and the rest *stx1a/stx2a*. During the later period, six local persistent lineages continued to be associated with human illness, including 74.7% (95% CI 68.3%, 80.3%) of reported cases in 2018 and 2019. The *stx* profile of isolates in local persistent lineages shifted from the earlier period, with 51.2% encoding only *stx2a*.

51 *CONCLUSIONS*

52 Our study identified multiple locally evolving lineages transmitted between cattle and humans
53 persistently associated with *E. coli* O157:H7 illnesses for up to 13 years. Of concern, there was a
54 dramatic shift in the local persistent lineages toward strains with the more virulent *stx2a*-only
55 profile. We hypothesize that the large proportion of disease associated with local transmission
56 systems is a principal cause of Alberta's high *E. coli* O157:H7 incidence.

57

58 **INTRODUCTION**

59 Several areas around the globe experience exceptionally high incidence of Shiga toxin-producing
60 *Escherichia coli* (STEC), including the virulent serotype *E. coli* O157:H7. These include
61 Scotland,[1] Ireland,[2] Argentina,[3] and the Canadian province of Alberta.[4] All are home to large
62 populations of agricultural ruminants, STEC's primary reservoir. However, there are many
63 regions with similar ruminant populations where STEC incidence is unremarkable. What
64 differentiates high risk regions is unclear. Moreover, with systematic STEC surveillance only
65 conducted in limited parts of the world,[5] there may be unidentified regions with exceptionally
66 high disease burden.
67
68 STEC infections can arise from local reservoirs, transmitted through food, water, direct animal
69 contact, or contact with contaminated environmental matrices. The most common reservoirs
70 include domesticated ruminants such as cattle, sheep, and goats. While STEC has been isolated
71 from a variety of other animal species and outbreaks have been linked to species such as deer[6]
72 and swine,[7] it is unclear what roles they play as maintenance or intermediate hosts. STEC
73 infections can be imported through food items traded nationally and internationally, as has been
74 seen with *E. coli* O157:H7 outbreaks in romaine lettuce from the United States.[8] Secondary
75 transmission is believed to cause approximately 15% of cases, but the pathogen is not believed to
76 be sustained long-term through person-to-person transmission alone.[9,10]
77
78 The mix of STEC infection sources in a region directly influences public health measures needed
79 to control disease burden. Living near cattle and other domesticated ruminants has been linked to
80 STEC incidence, particularly for *E. coli* O157:H7.[2,11-15] These studies suggest an important role

81 for local reservoirs in STEC epidemiology. A comprehensive understanding of STEC's disease
82 ecology would enable more effective investigations into potential local transmission systems and
83 ultimately their control. Here, we take a phylodynamic, genomic epidemiology approach to more
84 precisely discern the role of the cattle reservoir in the dynamics of *E. coli* O157:H7 human
85 infections. We focus on the high incidence region of Alberta, Canada to provide insight into
86 characteristics that make the pathogen particularly prominent in such regions.
87

88 **METHODS**

89 *STUDY DESIGN AND POPULATION*

90 We conducted a multi-host genomic epidemiology study in Alberta, Canada. Our primary
91 analysis focused on 2007 to 2015 due to the availability of isolates from intensive provincial
92 cattle studies.[16-19] To select both cattle and human isolates, we block randomized by year to
93 ensure representation across the period. We define isolates as single bacterial species obtained
94 from culture. We sampled 123 *E. coli* O157 cattle isolates from 4,660 available. Selected cattle
95 isolates represented 7 of 12 cattle study sites and 56 of 89 sampling occasions from the source
96 studies.[16-19] Samples were taken from fecal pats, rectal grabs, and hide swabs from cattle in
97 feedlots and fecal samples from transport trucks. We sampled 123 of 1,148 *E. coli* O157 isolates
98 collected from cases reported to the provincial health authority (Alberta Health) during the
99 corresponding time period (Supplemental Information).

100

101 In addition to the 246 isolates for the primary analysis, we contextualized our findings with three
102 additional sets of *E. coli* O157:H7 isolates (Figure 1): 445 from Alberta Health from 2009 to
103 2019 and already sequenced as part of other public health activities; 152 from the U.S. from
104 1999 to 2016; and 54 from elsewhere around the world between 2007 and 2015. The additional
105 Alberta Health isolates were sequenced by the National Microbiology Laboratory (NML)-Public
106 Health Agency of Canada (Winnipeg, Manitoba, Canada) as part of PulseNet Canada activities.
107 Isolates sequenced by the NML for 2018 and 2019 constituted the majority of reported *E. coli*
108 O157:H7 cases for those years (217 of 247; 87.9%). U.S. isolates were considered separately
109 from other global isolates, as the U.S. is Alberta's most frequent international trade partner, with
110 both processed beef and live cattle crossing the border. U.S. isolates from 1999 to 2009 and

111 global isolates were identified from previous literature,[20] and U.S. isolates from 2010 to 2016
112 were randomly selected from *E. coli* O157:H7 sequences available through the U.S. CDC's
113 PulseNet BioProject PRJNA218110.
114
115 This study was approved by the University of Calgary Conjoint Health Research Ethics Board,
116 #REB19-0510. A waiver of consent was granted, and all case data were deidentified.
117

118 *WHOLE GENOME SEQUENCING, ASSEMBLY, AND INITIAL PHYLOGENY*

119 The 246 isolates for the primary analysis were sequenced using Illumina NovaSeq 6000 and
120 assembled into contigs using the Unicycler v04.9 pipeline, as described previously (BioProject
121 PRJNA870153).[21] Raw read FASTQ files were obtained from Alberta Health for the additional
122 445 isolates sequenced by the NML and from NCBI for the 152 U.S. and 54 global sequences.
123 We used the SRA Toolkit v3.0.0 to download sequences for U.S. and global isolates using their
124 BioSample (i.e. SAMN) numbers. The corresponding FASTQ files could not be obtained for 6
125 U.S. and 7 global isolates we had selected (Figure 1).
126

127 PopPUNK v2.5.0 was used to cluster Alberta isolates and identify any outside the O157:H7
128 genomic cluster (Supplemental Figure S1).[22] For assembling and quality checking (QC) all
129 sequences, we used the Bactopia v3.0.0 pipeline.[23] This pipeline performed an initial QC step on
130 the reads using FastQC v0.12.1, which trimmed adapters and read ends with quality lower than 6
131 and discarded reads after trimming with overall quality scores lower than 10. None of the isolates
132 were eliminated during this step for low read quality. We used the Shovill v1.1.0 assembler
133 within the Bactopia pipeline to *de novo* assemble the Unicycler contigs for the primary analysis

134 and raw reads from the supplementary datasets. Bactopia generated a quality report on the
135 assemblies, which we assessed based on number of contigs (<500), genome size (≥5.1 Mb), N50
136 (>30,000), and L50 (≤50). Low-quality assemblies were removed. This included 1 U.S.
137 sequence, for which 2 FASTQ files had been attached to a single BioSample identifier; the other
138 sequence for the isolate passed all quality checks and remained in the analysis. Additionally, 16
139 sequences from the primary analysis dataset and 4 from the extended Alberta data had a total
140 length <5.1 Mb. These sequences corresponded exactly to those identified by the PopPUNK
141 analysis to be outside the primary *E. coli* O157:H7 genomic cluster. Finally, although all isolates
142 were believed to be of cattle or clinical origin during initial selection, detailed metadata review
143 identified 1 isolate of environmental origin in the primary analysis dataset and 8 that had been
144 isolated from food items in the extended Alberta data. These were excluded. We used
145 STECFinder v1.1.0[24] to determine Shiga toxin gene (*stx*) profile and confirm the *E. coli*
146 O157:H7 serotype using the *wzy* or *wzx* O157 O-antigen genes and detection of the H7 H-
147 antigen. After processing, we had 229 isolates (121 human, 108 cattle) in our primary sample,
148 432 additional Alberta Health isolates, 146 U.S. isolates, and 47 global isolates (Figure 1,
149 Supplemental Data File).
150
151 Bactopia's Snippy workflow, which incorporates Snippy v4.6.0, Gubbins v3.3.0, and IQTree
152 v2.2.2.7, followed by SNP-Sites v2.5.1, were used to generate a core genome SNP alignment
153 with recombinant blocks removed. The maximum likelihood phylogeny of the core genome SNP
154 alignment generated by IQTree was visualized in Microreact v251. The number of core SNPs
155 between isolates was calculated using PairSNP v0.3.1. Clade was determined based on the
156 presence of at least one defining SNP for the clade as published previously.[25] Isolates were

157 identified to the subclade level [e.g. G(vi)] when both clade and subclade SNPs were present and
158 the clade level (e.g. G) when only clade SNPs were present.
159

160 *PHYLODYNAMIC AND STATISTICAL ANALYSES*

161 For our primary analysis, we created a timed phylogeny, a phylogenetic tree on the scale of time,
162 in BEAST2 v2.6.7 using the structured coalescent model in the Mascot v3.0.0 package with
163 demes for cattle and humans (Supplemental Table S1). The analysis was run using four different
164 seeds to confirm that all converged to the same solution, and tree files were combined before
165 generating a maximum clade credibility (MCC) tree. State transitions between cattle and human
166 isolates over the entirety of the tree, with their 95% highest posterior density (HPD) intervals,
167 were also calculated from the combined tree files. We determined the influence of the prior
168 assumptions on the analysis (Supplemental Table S1) with a run that sampled from the prior
169 distribution (Supplemental Figure S2, Supplemental Information).
170

171 Local persistent lineages (LPLs) were identified based on following criteria: 1) a single lineage
172 of the MCC tree with a most recent common ancestor (MRCA) with ≥95% posterior probability;
173 2) all isolates <30 core SNPs from one another; 3) contained at least 1 cattle isolate; 4) contained
174 ≥5 isolates; and 5) the isolates were collected at sampling events (for cattle) or reported (for
175 humans) over a period of at least 1 year. From non-LPL isolates, we estimated the number of
176 local transient isolates vs. imported isolates. For the 121 human *E. coli* O157:H7 isolates in the
177 primary sample, we determined what portion belonged to local persistent lineages (LPL) and
178 what portion were likely to be from local transient *E. coli* O157:H7 populations vs. imported.
179 Human isolates within the LPLs were enumerated (n=44). The 77 human isolates outside LPLs

180 included 58 clade G(vi) isolates and 19 non-G(vi) isolates. Based on the MCC tree from the
181 primary analysis, none of the non-G(vi) human isolates was likely to have been closely related to
182 an isolate from the Alberta cattle population, suggesting that all 19 were imported. As a
183 proportion of all non-LPL human isolates, these 19 constituted 24.7%. While it may be possible
184 that all clade G(vi) isolates were part of a local evolving lineage, it is also possible that exchange
185 of both cattle and food from other locations was causing the regular importation of clade G(vi)
186 strains and infections. Thus, we used the proportion of non-LPL human isolates outside the G(vi)
187 clade to estimate the proportion of non-LPL human isolates within the G(vi) clade that were
188 imported; i.e., $58 \times 24.7\% = 14$. We then conducted a similar exercise for cattle isolates.
189
190 To contextualize our results in terms of ongoing human disease burden, we created a timed
191 phylogeny using a constant, unstructured coalescent model of the 229 Alberta isolates from the
192 primary analysis and the additional Alberta Health isolates. Outbreaks were down-sampled to
193 avoid biasing the tree by randomly selecting 1 to 2 isolates per outbreak; as such, only 230 of the
194 432 additional isolates were included in the analysis (Supplemental Table S1). We identified
195 LPLs as above, and leveraged the near-complete sequencing of isolates from 2018 and 2019 to
196 calculate the proportion of reported human cases associated with LPLs.
197
198 We created a timed phylogeny of Alberta isolates and U.S. isolates from 1996 to 2016 to test
199 whether the LPLs or Alberta's dominant *E. coli* O157:H7 clade (G) were linked to U.S. ancestors
200 (Supplemental Table S1). We also created a timed phylogeny of temporally overlapping Alberta,
201 U.S., and global isolates from 2007 to 2015, excluding clades A and B, which were too limited to
202 make meaningful comparisons.

All BEAST2 analyses were run for 100,000,000 Markov chain Monte Carlo iterations or until all parameters converged with effective sample sizes >200, whichever was longer. Exact binomial 95% confidence intervals (CIs) were computed for proportions.

208 **RESULTS**

209 Across the 854 isolates included in the analyses, we identified 11,234 core genome SNPs. The
210 monophyletic clade G(vi) constituted 74.4% (n=635) of all isolates (Figure 2). The majority of
211 all Alberta isolates belonged to the G(vi) clade (582 of 661; 88.0%), compared to 51 (34.9%) of
212 the U.S. isolates and 2 (4.3%) of the global isolates (Table 1). There were 487 (76.7%) clade
213 G(vi) isolates with the *stx1a/stx2a* profile, compared to 1 (0.5%) among the 219 isolates outside
214 the G(vi) clade.

215

216 *THE MAJORITY OF CLINICAL CASES EVOLVED FROM LOCAL CATTLE LINEAGES*

217 In our primary sample of 121 human and 108 cattle isolates from Alberta from 2007 to 2015,
218 SNP distances were comparable between species (Figure 3a). Among sampled human cases, 19
219 (15.7%; 95% CI 9.7%, 23.4%) were within 5 SNPs of a sampled cattle strain.

220

221 The phylogeny generated by our primary structured coalescent analysis indicated cattle were the
222 primary reservoir, with a high probability that the hosts at nodes along the tree's backbone were
223 cattle (Figure 3b). The root was estimated at 1812 (95% HPD 1748, 1870). The most recent
224 common ancestor (MRCA) of clade G(vi) strains in Alberta was inferred to be a cattle strain,
225 dated to 1971 (95% HPD 1961, 1980). With our assumption of a relaxed molecular clock, the
226 mean clock rate for the core genome was estimated at 1.00×10^{-4} (95% HPD 8.45×10^{-5}, 1.18×10^{-4})
227 substitutions/site/year. The effective population size, N_e, of the human *E. coli* O157:H7
228 population was estimated as 913 (95% HPD 620, 1232), and for cattle as 49 (95% HPD 32, 67).
229 We estimated 107 (95% HPD 101, 111) human lineages arose from cattle lineages, and 31 (95%
230 HPD 22, 43) arose from other human lineages (Figure 3c). In other words, 77.5% of human

231 lineages arose from cattle lineages. We observed minimal influence of our choice of priors
232 (Supplemental Figure S2, Supplemental Text).
233

234 *Local Persistent Lineages Account for the Majority of Ongoing Human Disease*

235 In our primary analysis, we identified 11 local persistent lineages (LPLs) (Figure 3b). LPLs
236 included a range of 5 (G(vi)-AB LPLs 9 and 10) to 26 isolates (G(vi)-AB LPL 2), with an
237 average of 10. LPLs tended to be clustered on the MCC tree. G(vi)-AB LPLs 1-4, 6-8, and 9 and
238 10 were clustered with MRCAs inferred at 1997 (95% HPD 1993, 2000), 1998 (95% HPD 1995,
239 2001), and 1996 (95% HPD 1993, 1999), respectively. Cattle were the inferred host of all three
240 ancestral nodes.
241

242 LPLs included 71 of 108 (65.7%; 95% CI 56.0%, 74.6%) cattle and 44 of 121 (36.4%; 95% CI
243 27.8%, 45.6%) human isolates. Of the remaining human isolates, 33 (27.3%) were associated
244 with imported infections and 44 (36.4%) with infections from transient local strains. Of the
245 remaining cattle isolates, 11 (10.2%) were imported and 26 (24.1%) were associated with
246 transmission from transient strains. Of the 115 isolates in LPLs, 7 (6.1%) carried only *stx2a*, and
247 the rest *stx1a/stx2a*. Among the 114 non-LPL isolates, 27 (23.7%) were *stx2a*-only, 1 (0.9%) was
248 *stx1a*-only, 6 (5.3%) were *stx1a/stx2c*, and the remaining 80 (70.2%) were *stx1a/stx2a*.
249

250 To understand long-term persistence, we expanded the phylogeny with additional Alberta Health
251 isolates from 2009 to 2019 (Supplemental Table S1). Six of the 11 LPLs identified in our
252 primary analysis continued to cause disease during the 2016 to 2019 period (Figure 4a). With
253 most of the cases reported during 2018 and 2019 sequenced, we were able to estimate the

254 proportion of reported *E. coli* O157:H7 associated with LPLs. Of 217 sequenced cases reported
255 during these two years, 162 (74.7%; 95% CI 68.3%, 80.3%) arose from Alberta's LPLs. The *stx*
256 profile of LPL isolates shifted as compared to the primary analysis, with 83 (51.2%) of the LPL
257 isolates encoding only *stx2a* and the rest *stx1a/stx2a* (Figure 4b). Among the 55 non-LPL
258 isolates during 2018-2019, the *stx2c*-only profile emerged with 16 (29.1%) isolates, and *stx2a*-
259 only was found in only 6 (10.9%) cases.
260
261 All 5 large (≥10 cases) sequenced outbreaks in Alberta during the study period were within clade
262 G(vi). LPLs gave rise to 3 large outbreaks, accounting for 117 cases, including 83 from an
263 extended outbreak by a single strain, defined as isolates within 5 SNPs of one another, during
264 2018 and 2019 (Figure 4a). The two large outbreaks that did not arise from LPLs both occurred
265 in 2014 and were responsible for 164 cases.
266
267 *International Importation Does Not Explain Alberta's Current Disease Burden*
268 Only 2 U.S. isolates coincided with Alberta LPLs, specifically G(vi)-AB LPL 9 in 2014 and
269 G(vi)-AB LPL 11 in 2015 (Supplemental Figure S3). Isolates in these two LPLs from Alberta
270 dated to 2007 and 2009, respectively, and were identified multiple times up to and including
271 during the 2018-2019 period (Figure 4a). There was no evidence of early U.S. ancestors of LPLs.
272 No LPL contained a global isolate. Based on migration events calculated from the tree, we
273 estimated that 15.4% of combined human and cattle Alberta lineages were imported
274 (Supplemental Table S2). Sequences from outside North America were separated from Alberta
275 sequences by a median of 325 (IQR 288-349) SNPs. Including U.S. and global isolates in the

276 phylogeny did not alter the LPLs identified, though some minor rearrangement of the tree was

277 observed (Supplemental Figure S3).

278

279 **DISCUSSION**

280 Focusing on a region that experiences an especially high incidence of STEC, we conducted a
281 deep genomic epidemiologic analysis of *E. coli* O157:H7's multi-host disease dynamics. Our
282 study identified multiple locally evolving lineages transmitted between cattle and humans. These
283 were persistently associated with *E. coli* O157:H7 illnesses over periods of up to 13 years. Of
284 clinical importance, there was a dramatic shift in the *stx* profile of the strains arising from local
285 persistent lineages toward strains carrying only *stx2a*, which has been associated with increased
286 progression to hemolytic uremic syndrome (HUS).[26] We hypothesize that the large proportion of
287 cases associated with local transmission systems is a principal cause of Alberta's high *E. coli*
288 O157:H7 incidence.

289

290 Our study has provided quantitative estimates of cattle-to-human migration in a high incidence
291 region, the first such estimates of which we are aware. Our estimates are consistent with prior
292 work that established an increased risk of STEC associated with living near cattle.[2,11-15] We
293 showed that 77% of strains infecting humans arose from cattle lineages. These transitions can be
294 seen as a combination of the historic evolution of *E. coli* O157:H7 from cattle in the rare clades
295 and the infection of humans from local cattle or cattle-related reservoirs in clade G(vi). While
296 our findings indicate the majority of human cases arose from cattle lineages, transmission may
297 have involved intermediate hosts or environmental reservoirs several steps removed from the
298 cattle reservoir. However, our analysis demonstrates that local cattle remain an integral part of
299 the transmission system for the vast majority of cases, even when they may not be the immediate
300 source of infection.

301

302 The cattle-human transitions we estimated were based on structured coalescent theory,[27] which
303 we used throughout our analyses. This approach is similar to phylogeographic methods that have
304 previously been applied to *E. coli* O157:H7.[20] We inferred the full backbone of the Alberta *E.*
305 *coli* O157:H7 phylogeny as arising from cattle, consistent with the postulated global spread of
306 the pathogen via ruminants.[20] Our estimate of the origin of the serotype, at 1812 (95% HPD
307 1748, 1870), was somewhat earlier than previous estimates, but consistent with global (1890;
308 95% HPD 1845, 1925)[20] and United Kingdom (1840; 95% HPD 1817, 1855)[28] studies that used
309 comparable methods. Our dating of Alberta's G(vi) clade to 1971 (95% HPD 1961, 1980) also
310 corresponds to proposed migrations of clade G into Canada from the U.S. in 1965-1977.[20] Our
311 study thus adds to the growing body of work on the larger history of *E. coli* O157:H7, providing
312 an in-depth examination of the G(vi) clade.
313
314 Our identification of the 11 local persistent lineages (LPLs) is significant in demonstrating that
315 the majority of Alberta's reported *E. coli* O157:H7 illnesses are of local origin. Our definition
316 ensured that every LPL had an Alberta cattle strain and at least 5 isolates separated by >1 year,
317 making the importation of the isolates in a lineage highly unlikely. Further supporting the
318 evolution of the LPLs within Alberta, all 11 LPLs were in clade G(vi), several were
319 phylogenetically related with MRCAs dating to the late 1990s, and few non-Alberta isolates fell
320 within LPLs. The two U.S. isolates associated with Alberta LPLs may reflect Alberta cattle that
321 were slaughtered in the U.S. Thus, we are confident that the identified LPLs represent locally
322 evolving lineages and potential persistent sources of disease.
323

324 Based on our LPL analysis, we estimated only 27% of human and 10% of cattle *E. coli* O157:H7
325 isolates were imported. This was consistent with the overall importation estimate of 15% for all
326 Alberta lineages from our global structured coalescent analysis. While these estimates may
327 appear low given the recent focus on row crops and other produce as potential vehicles of
328 disease,[8] 26% of sporadic STEC infections have been attributed to animal contact and the farm
329 environment, with a further 19% to pink or raw meat.[10] Similarly, 24% of *E. coli* O157 outbreaks
330 in the U.S. were attributed to beef, animal contact, water, or other environmental sources.[9] In
331 Alberta, these are all inherently local exposures, given that 90% of beef consumed in Alberta is
332 produced and/or processed there. Even person-to-person transmission, responsible for 15% of
333 sporadic cases and 16% of outbreaks,[9,10] includes secondary transmission from cases infected
334 from local sources, which may explain our estimate of 23% for person-to-person transmission.
335 To our knowledge, our study provides the first comprehensive determination of local vs.
336 imported status for *E. coli* O157:H7 cases. Similar studies in regions of both high and moderate
337 incidence would provide further insight into the role of localization on *E. coli* O157:H7
338 incidence.
339
340 Of the 11 lineages we identified as LPLs during the 2007-2015 period, 6 were also identified in
341 the 2016-2019 period. During the initial period, 36% of human cases were linked to an LPL, and
342 6.1% carried only *stx2a*. The risk of HUS increases in strains of STEC carrying only *stx2a*,
343 relative to *stx1a/stx2a*,[26] meaning the earlier LPL population had few of the high-virulence
344 strains. In 2018 and 2019, the 6 long-term LPLs were associated with both greater incidence and
345 greater virulence, encompassing 75% of human cases with more than half of LPL isolates
346 carrying only *stx2a*. The cause of this shift remains unclear, though shifts toward greater

347 virulence in *E. coli* O157:H7 populations have been seen elsewhere.[29] The growth and diversity
348 of G(vi)-AB LPL 1 and G(vi)-AB LPL 6 in the later period suggest these lineages were in stable
349 reservoirs or adapted easily to new niches. Identifying these reservoirs could yield substantial
350 insights into disease prevention, given the significant portion of illnesses caused by persistent
351 strains.
352
353 We developed a novel measure of persistence for use in this study, specifically for the purposes
354 of identifying lineages that pose an ongoing threat to public health in a specific region.
355 Persistence has been defined variably in the literature, for example as shedding of the same strain
356 for at least 4 months.[30] Most recently, the U.S. CDC has identified the first Recurring, Emergent,
357 and Persistent (REP) STEC strain, REPEXH01, which has been detected since 2017 in over 600
358 cases. REPEXH01strains are within 21 allele differences of one another
359 (https://www.cdc.gov/ncezid/dfwed/outbreak-response/rep-strains/repexh01.html). Given that we
360 used high resolution SNP analysis rather than MLST, we used a difference of <30 SNPs to
361 define persistent lineages. Supporting the persistence we have observed, the REPEXH01 strain is
362 also an *E. coli* O157:H7 strain; however, O157:H7 was defined as sporadic in a German study
363 using the 4-month shedding definition, which may be due to ecological differences.[30]
364 Understanding microbial drivers of persistence is an active field of research, with early findings
365 suggesting a correlation of STEC persistence to the accessory genome and traits such as biofilm
366 formation and nutrient metabolism.[30,31] Our approach to studying persistence was specifically
367 designed for longitudinal sampling in high-incidence regions and may be useful for others
368 attempting to identify sources that disproportionately contribute to disease burden.
369

370 Our analysis was limited to only cattle and humans. However, small ruminants (e.g., sheep,
371 goats) have also been identified as important STEC reservoirs,[12,15,25] and Alberta has experienced
372 outbreaks linked to swine.[7] Had isolates from a wider range of potential reservoirs been
373 available, we would have been able to elucidate more clearly the roles that various hosts and
374 common sources of infection play in local transmission. This may help explain the 3 human-to-
375 cattle predicted transmissions, which could be erroneous. We also limited our analysis only to *E.*
376 *coli* O157:H7 despite the growing importance of non-O157 STEC as historical multi-species
377 collections of non-O157 isolates are lacking. As serogroups differ meaningfully in exposures,[32]
378 our results may not be generalizable beyond the O157 serogroup. Finally, we were not able to
379 estimate the impact of strain migration between Alberta and the rest of Canada, because
380 metadata for publicly-available *E. coli* O157:H7 sequences from Canada was limited, such that
381 we could not be sure they were from outside Alberta.
382
383 *E. coli* O157:H7 infections are a pressing public health problem in many high incidence regions
384 around the world including Alberta, where a recent childcare outbreak caused >300 illnesses. In
385 the majority of sporadic cases, and even many outbreaks,[9] the source of infection is unknown,
386 making it critical to understand the disease ecology of *E. coli* O157:H7 at a system level. Here
387 we have identified a high proportion of human cases arising from cattle lineages and a low
388 proportion of imported cases. Local transmission systems, including intermediate hosts and
389 environmental reservoirs, need to be elucidated to develop management strategies that reduce the
390 risk of STEC infection. In Alberta, local transmission is dominated by a single clade, and over
391 the extended study period, persistent lineages caused an increasing proportion of disease. The
392 local lineages with long-term persistence are of particular concern because of their increasing

393 virulence, yet they also present opportunity as larger, more stable targets for reservoir

394 identification and control.

395

ACKNOWLEDGEMENTS

We would like to acknowledge Dr. Angela Ma, Hannah Tyrrell, and Dr. Surangi Thilakarathna for their work preparing clinical isolates for sequencing, and Dr. Jesse Berman for reviewing an early version of this manuscript.

FUNDING

Funding for this work was provided by the Beef Cattle Research Council (FOS.01.18). The sponsor had no role in the study design; collection, analysis, or interpretation of data; writing of the report; or the decision to submit the paper for publication.

DECLARATION OF INTERESTS

The authors declare no conflicts of interest.

AUTHOR CONTRIBUTIONS

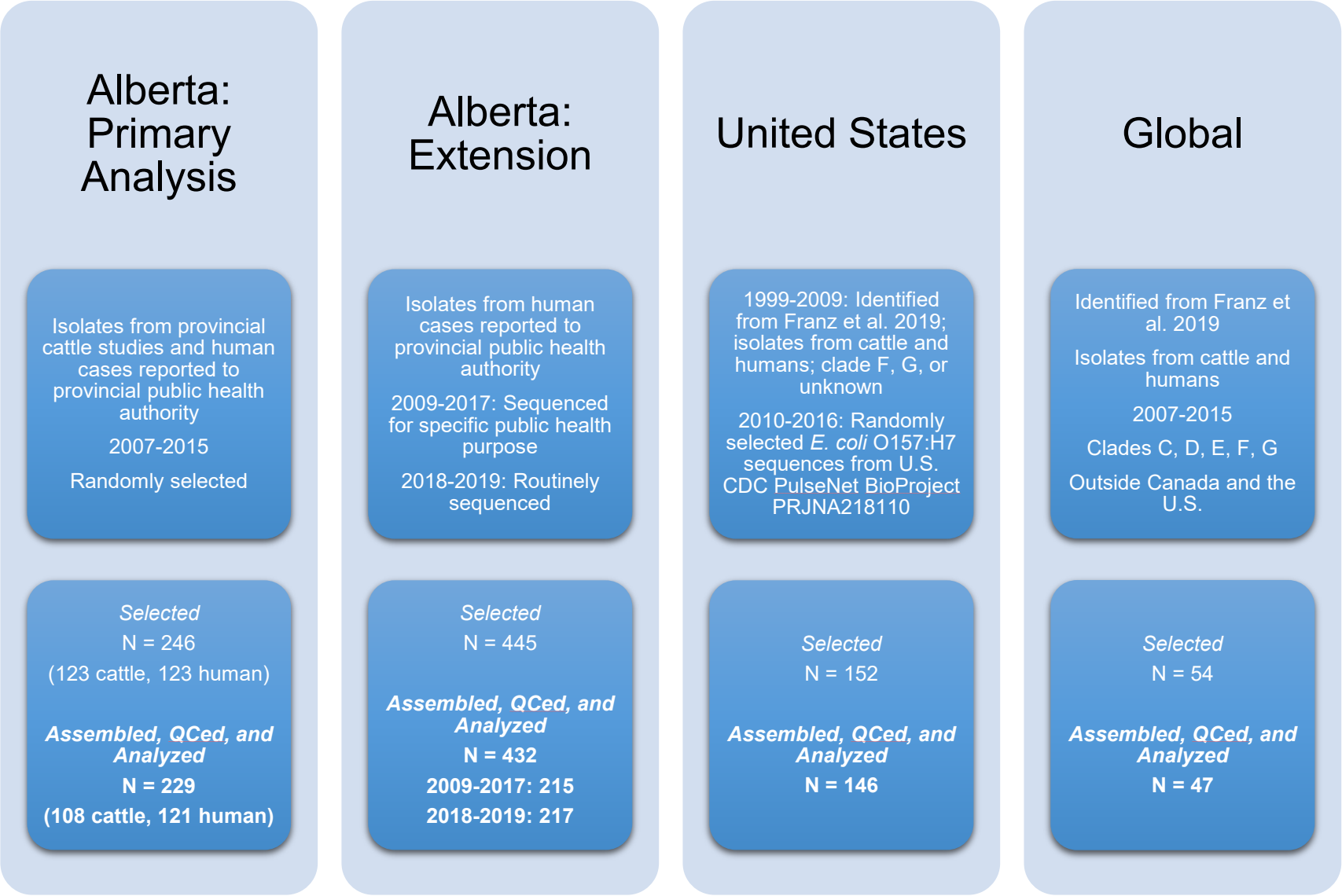
Gillian A.M. Tarr: Conceptualization, Methodology, Software, Formal analysis, Data curation, Writing - Original Draft, Visualization, Funding acquisition. **Linda Chui**: Conceptualization, Methodology, Resources, Writing - Review & Editing, Supervision, Funding acquisition. **Kim Stanford**: Conceptualization, Methodology, Resources, Writing - Review & Editing, Supervision, Funding acquisition. **Emmanuel W. Bumunang**: Investigation, Data curation, Writing - Review & Editing. **Rahat Zaheer**: Methodology, Investigation, Writing - Review & Editing. **Vincent Li**: Validation, Data curation, Writing - Review & Editing. **Stephen B. Freedman**: Conceptualization, Writing - Review & Editing, Project administration, Funding acquisition. **Chad Laing**: Conceptualization, Methodology, Writing - Review & Editing,

Funding acquisition. **Tim A. McAllister**: Conceptualization, Methodology, Resources, Writing - Review & Editing, Supervision, Project administration, Funding acquisition.

DATA SHARING STATEMENT

Data from this study not previously published will be made available at publication. Deidentified participant data, associated NCBI accession numbers for sequence data, and an accompanying data dictionary will be provided as an attached data supplement.

427 **FIGURES**

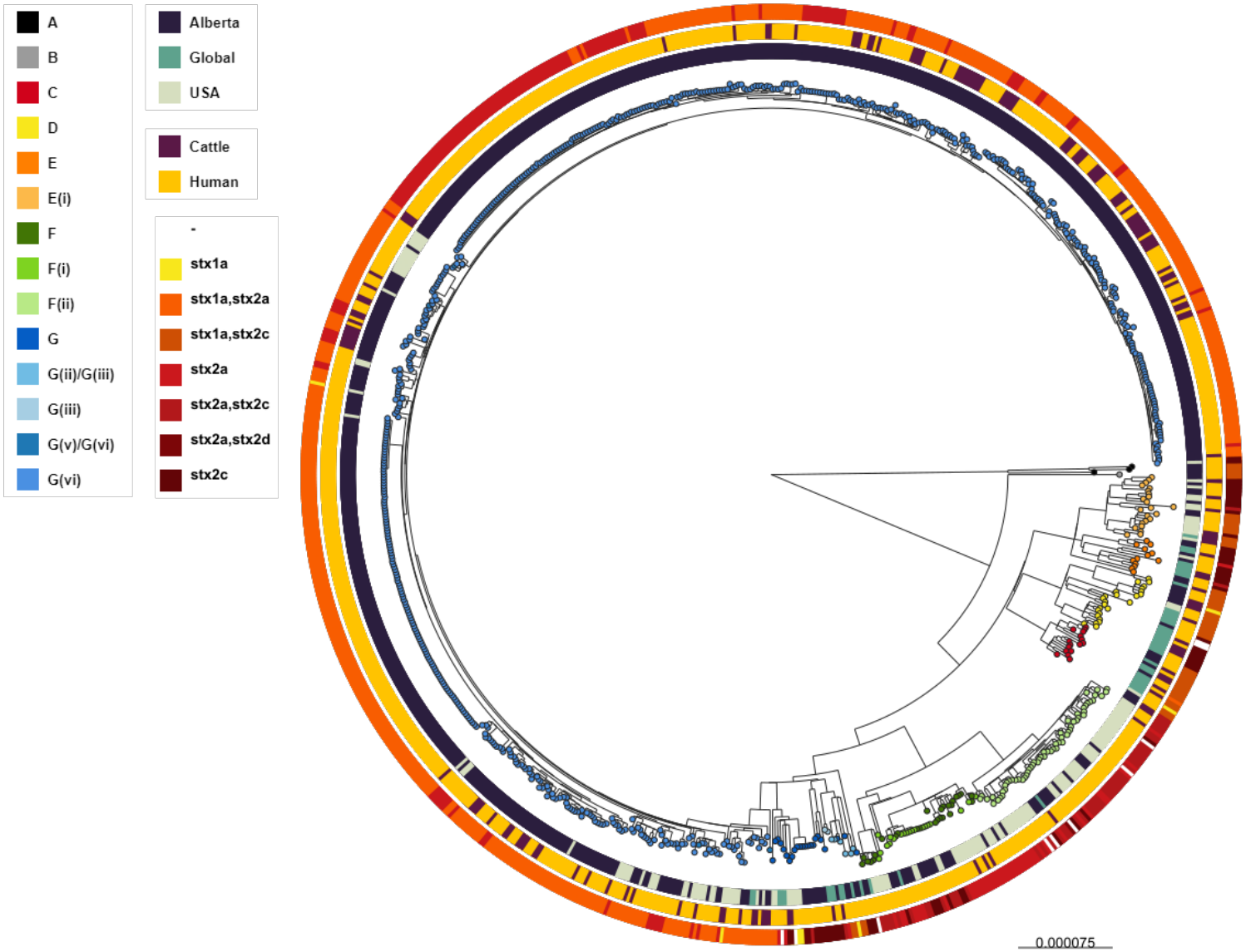


428

429 **Figure 1. *E. coli* O157:H7 isolates selected for the study and included in analysis.** Four sets of isolates, all originating from cattle

430 or humans, were included in the study.

431


A
B
C
D
E
E(i)
F
F(i)
F(ii)
G
G(ii)/G(iii)
G(iii)
G(v)/G(vi)
G(vi)
Alberta
Global
USA
Cattle
Human
-
stx1a
stx1a,stx2a
stx1a,stx2c
stx2a
stx2a,stx2c
stx2a,stx2d
stx2c
0.000075

432

433 **Figure 2. Maximum likelihood core SNP tree of the 854 *E. coli* O157:H7 isolates referenced in the study.** This includes 229
434 randomly sampled cattle and human isolates from Alberta, Canada, from 2007 through 2015; 432 additional Alberta isolates
435 sequenced as part of public health activities from 2009 through 2019; and 152 isolates from the U.S. and 47 isolates from elsewhere
436 around the globe from 1996 to 2016 to examine international transmission history. Clade is shown in the coloration of the tips on the
437 tree, geographic origin is shown on the inner ring, species of origin on the middle ring, and Shiga toxin gene (*stx*) profile on the outer
438 ring. The tree was rooted at clade A. Clade G constituted the majority of isolates.
439

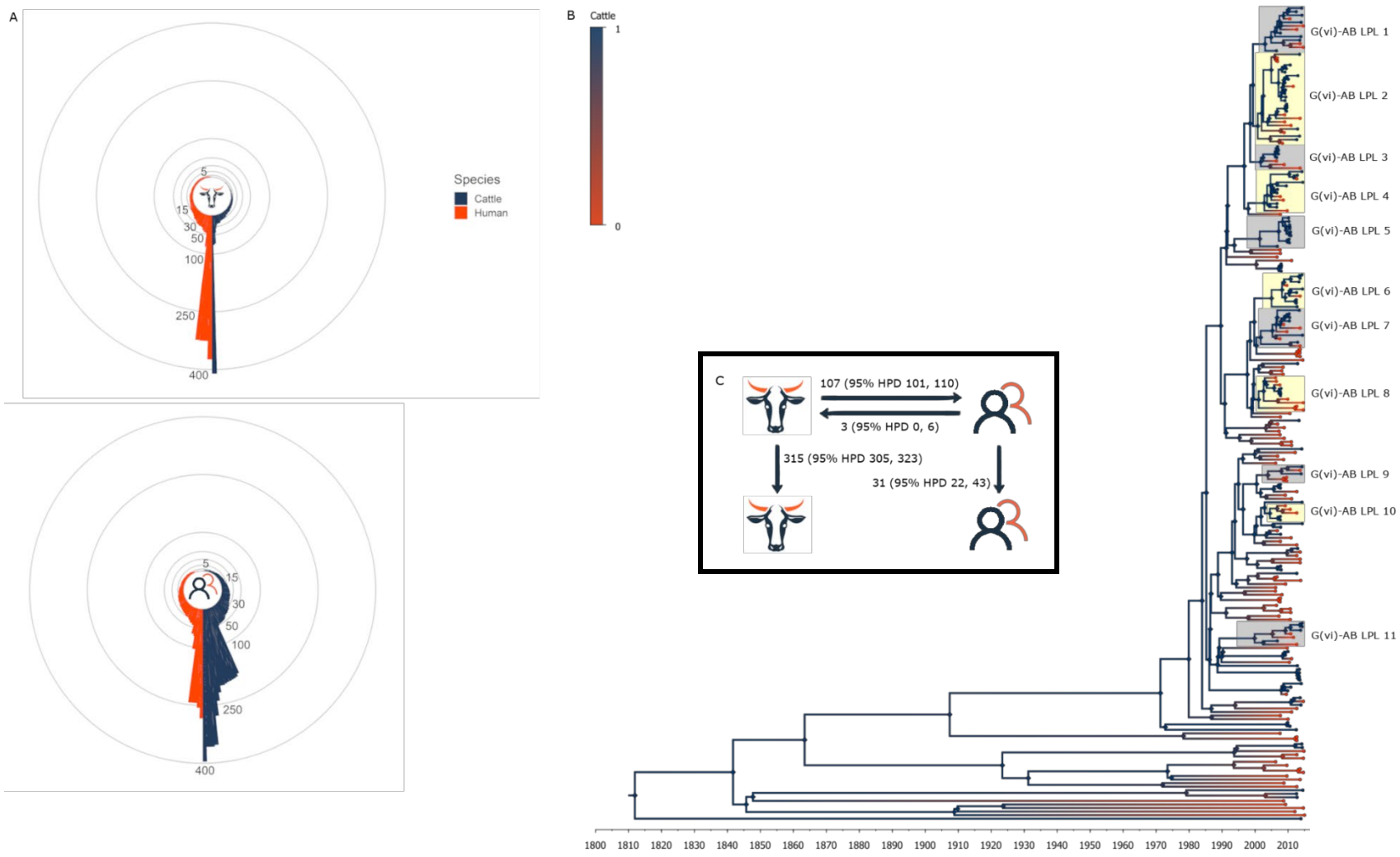


Figure 3. Relationship of randomly selected *E. coli* O157:H7 strains isolated from 121 reported human cases and 108 beef cattle in Alberta, Canada, 2007-2015. Target diagrams show SNP distances from cattle (A, top) and humans (A, bottom) to cattle

443 (blue) and humans (orange), with rings labeled with the SNP distance between isolates. Cattle isolates were highly related with 53% of
444 cattle isolates within 5 SNPs of another cattle isolate and 83% within 15 SNPs (A, top). Human isolates showed a bimodal distribution
445 in their relationship to cattle isolates, with 87% within 52 SNPs of a cattle isolate and the remainder 185-396 SNPs apart (A, bottom).
446 The maximum clade credibility tree for the structured coalescent analysis of cattle and human isolates (B) was colored by inferred
447 host, cattle (blue) or human (orange). The majority of ancestral nodes inferred as cattle suggests cattle as the primary reservoir. The
448 root was estimated at 1812 (95% HPD 1748, 1870). Eleven local persistent lineages (LPLs) were identified, all in the G(vi) clade and
449 labeled G(vi)-AB LPL 1 through 11 (yellow and gray coloration highlights LPLs but has no other meaning). These accounted for 44
450 human (36.4%) and 71 cattle (65.7%) isolates. The structured coalescent model estimated 107 cattle-to-human state transitions
451 between branches, compared to only 31 human-to-human transitions, inferring cattle as the origin of 77.5% of human lineages (C).
452

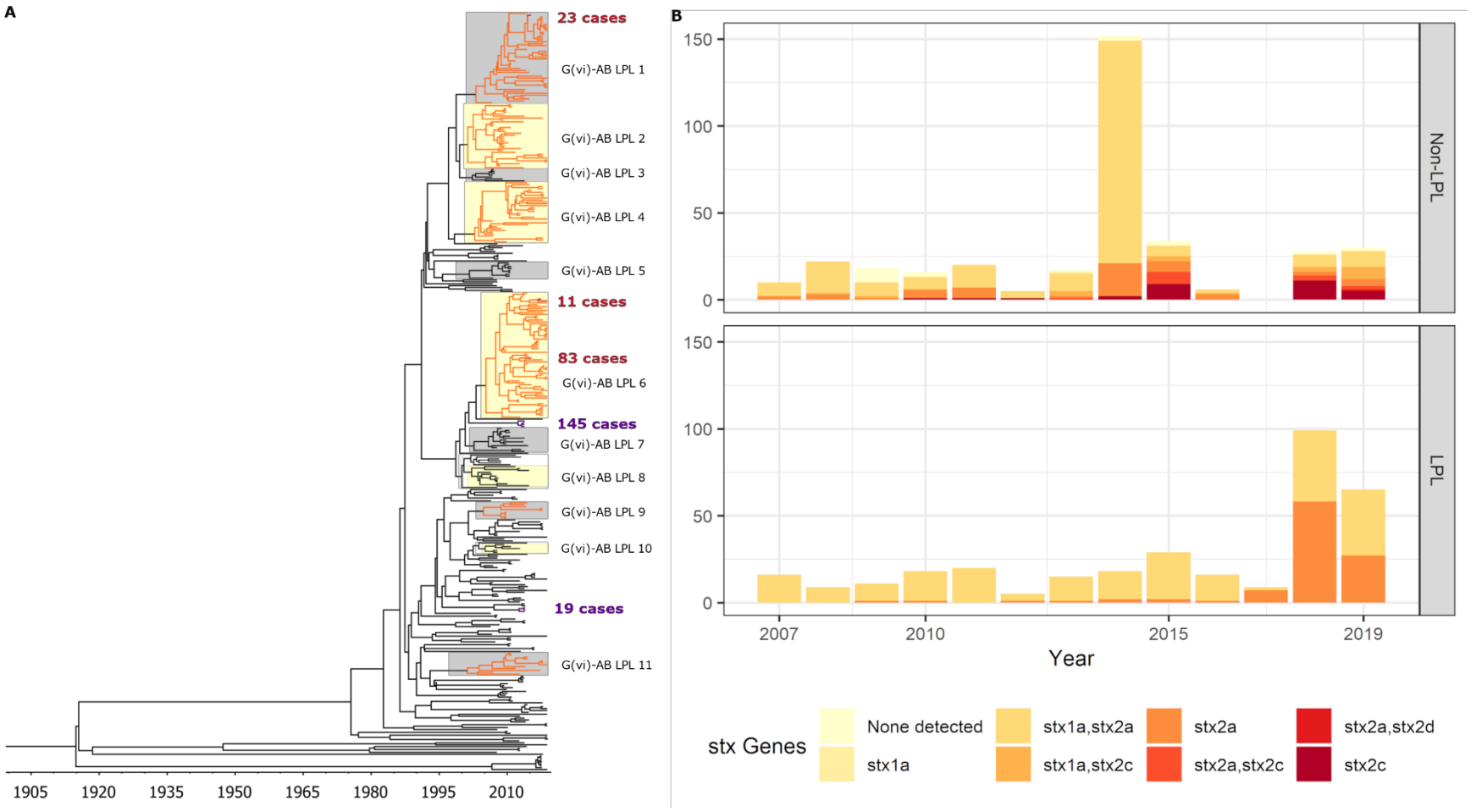


Figure 4. Extension of Alberta, Canada *E. coli* O157:H7 analysis to include 229 randomly selected study isolates and 432 additional public health isolates available from 2009 to 2019. Six local persistent lineages (LPLs) in clade G continued to be associated with disease after the initial study period, as indicated by branches colored in orange (A). LPLs are shaded and labeled as in Figure 3. Outbreaks reported during the period were down-sampled to avoid biasing the phylogeny, and the number of cases

458 represented associated with each outbreak are shown in red (LPL-associated outbreaks) or purple (non-LPL-associated outbreaks) text.
459 In 2018 and 2019, 74.7% of reported cases were associated with an LPL. The *stx* profile across all clades shifted from the initial study
460 period (2007-2015) to the later study period (2016-2019), with more of the virulent *stx2a*-only profile observed in 2018 and 2019 than
461 in previous years (B). The peak in sequences in 2014 is due to two outbreaks; routine sequencing began in 2018 and 2019, accounting
462 for the rise in sequenced cases during those years.
463

464 **TABLES**

465 **Table 1.** Distribution of study isolates by geographic source, clade, and Shiga toxin gene (*stx*) profile

Clade	Source	*stx1a*	*stx1a/stx2a*	*stx1a/stx2c*	*stx2a*	*stx2c*	*stx2a/stx2c*	*stx2a/stx2d*	None detected	Total
G(vi)		**2**	**487**	**0**	**146**	**0**	**0**	**0**	**0**	**635**
	Alberta	0	443	0	139	0	0	0	0	582
	U.S.	2	42	0	7	0	0	0	0	51
	Global	0	2	0	0	0	0	0	0	2
Other G		**3**	**0**	**4**	**3**	**16**	**2**	**0**	**3**	**31**
	Alberta	0	0	4	0	10	1	0	0	15
	U.S.	2	0	0	0	2	0	0	1	5
	Global	1	0	0	3	4	1	0	2	11
F		**0**	**0**	**0**	**41**	**11**	**49**	**1**	**5**	**107**
	Alberta	0	0	0	12	9	12	1	1	35
	U.S.	0	0	0	29	2	35	0	4	70
	Global	0	0	0	0	0	2	0	0	2
Other (A-E)		**2**	**1**	**38**	**1**	**35**	**2**	**0**	**2**	**81**
	Alberta	1	0	15	0	13	0	0	0	29
	U.S.	0	0	9	1	9	1	0	0	20
	Global	1	1	14	0	13	1	0	2	32
Total		**7**	**488**	**42**	**191**	**62**	**53**	**1**	**10**	**854**

466

467 **REFERENCES**

468 1. Innocent GT, Mellor DJ, McEwen SA, et al. Spatial and temporal epidemiology of
469 sporadic human cases of Escherichia coli O157 in Scotland, 1996-1999. *Epidemiol Infect* 2005;
470 **133**(6): 1033-41.

471 2. Ohaiseadha C, Hynds PD, Fallon UB, O'Dwyer J. A geostatistical investigation of
472 agricultural and infrastructural risk factors associated with primary verotoxigenic E. coli (VTEC)
473 infection in the Republic of Ireland, 2008-2013. *Epidemiol Infect* 2017; **145**(1): 95-105.

474 3. Rivero MA, Passucci JA, Rodriguez EM, Parma AE. Role and clinical course of
475 verotoxigenic Escherichia coli infections in childhood acute diarrhoea in Argentina. *J Med*
476 *Microbiol* 2010; **59**(Pt 3): 345-52.

477 4. Lisboa, Szelewicki, Lin, et al. Epidemiology of Shiga Toxin-Producing Escherichia coli
478 O157 in the Province of Alberta, Canada, 2009–2016. *Toxins* 2019; **11**(10): 613.

479 5. Kirk MD, Pires SM, Black RE, et al. World Health Organization Estimates of the Global
480 and Regional Disease Burden of 22 Foodborne Bacterial, Protozoal, and Viral Diseases, 2010: A
481 Data Synthesis. *PLoS Med* 2015; **12**(12): e1001921.

482 6. Laidler MR, Tourdjman M, Buser GL, et al. Escherichia coli O157:H7 Infections
483 Associated With Consumption of Locally Grown Strawberries Contaminated by Deer. *Clinical*
484 *Infectious Diseases* 2013; **57**(8): 1129-34.

485 7. Honish L, Punja N, Nunn S, et al. Escherichia coliO157:H7 Infections Associated
486 with Contaminated Pork Products — Alberta, Canada, July–October 2014. *MMWR Morbidity*
487 *and Mortality Weekly Report* 2017; **65**(52): 1477-81.

488 8. Bottichio L, Keaton A, Thomas D, et al. Shiga Toxin-Producing Escherichia coli
489 Infections Associated With Romaine Lettuce-United States, 2018. *Clin Infect Dis* 2020; **71**(8):
490 e323-e30.
491 9. Tack DM, Kisselburgh HM, Richardson LC, et al. Shiga Toxin-Producing Escherichia
492 coli Outbreaks in the United States, 2010-2017. *Microorganisms* 2021; **9**(7).
493 10. Kintz E, Brainard J, Hooper L, Hunter P. Transmission pathways for sporadic Shiga-
494 toxin producing E. coli infections: A systematic review and meta-analysis. *Int J Hyg Environ*
495 *Health* 2017; **220**(1): 57-67.
496 11. Jaros P, Cookson AL, Campbell DM, et al. A prospective case-control and molecular
497 epidemiological study of human cases of Shiga toxin-producing Escherichia coli in New
498 Zealand. *BMC Infect Dis* 2013; **13**: 450.
499 12. Elson R, Grace K, Vivancos R, et al. A spatial and temporal analysis of risk factors
500 associated with sporadic Shiga toxin-producing Escherichia coli O157 infection in England
501 between 2009 and 2015. *Epidemiol Infect* 2018; **146**(15): 1928-39.
502 13. Frank C, Kapfhammer S, Werber D, Stark K, Held L. Cattle density and Shiga toxin-
503 producing Escherichia coli infection in Germany: increased risk for most but not all serogroups.
504 *Vector Borne Zoonotic Dis* 2008; **8**(5): 635-43.
505 14. Friesema IH, Van De Kassteele J, CM DEJ, Heuvelink AE, Van Pelt W. Geographical
506 association between livestock density and human Shiga toxin-producing Escherichia coli O157
507 infections. *Epidemiol Infect* 2011; **139**(7): 1081-7.
508 15. Mulder AC, van de Kassteele J, Heederik D, Pijnacker R, Mughini-Gras L, Franz E.
509 Spatial Effects of Livestock Farming on Human Infections With Shiga Toxin-Producing

510 Escherichia coli O157 in Small but Densely Populated Regions: The Case of the Netherlands.
511 *Geohealth* 2020; **4**(11): e2020GH000276.
512 16. Stephens TP, McAllister TA, Stanford K. Perineal swabs reveal effect of super shedders
513 on the transmission of Escherichia coli O157:H7 in commercial feedlots. *J Anim Sci* 2009;
514 **87**(12): 4151-60.
515 17. Stanford K, Hannon S, Booker CW, Jim GK. Variable efficacy of a vaccine and direct-
516 fed microbial for controlling Escherichia coli O157:H7 in feces and on hides of feedlot cattle.
517 *Foodborne Pathog Dis* 2014; **11**(5): 379-87.
518 18. Stanford K, Gibb D, McAllister TA. Evaluation of a shelf-stable direct-fed microbial for
519 control of Escherichia coli O157 in commercial feedlot cattle. *Canadian Journal of Animal*
520 *Science* 2013; **93**(4): 535-42.
521 19. Stanford K, Johnson RP, Alexander TW, McAllister TA, Reuter T. Influence of Season
522 and Feedlot Location on Prevalence and Virulence Factors of Seven Serogroups of Escherichia
523 coli in Feces of Western-Canadian Slaughter Cattle. *PLoS One* 2016; **11**(8): e0159866.
524 20. Franz E, Rotariu O, Lopes BS, et al. Phylogeographic Analysis Reveals Multiple
525 International transmission Events Have Driven the Global Emergence of Escherichia coli
526 O157:H7. *Clin Infect Dis* 2019; **69**(3): 428-37.
527 21. Bumunang EW, Zaheer R, Stanford K, et al. Genomic Analysis of Shiga Toxin-
528 Producing E. coli O157 Cattle and Clinical Isolates from Alberta, Canada. *Toxins (Basel)* 2022;
529 **14**(9).
530 22. Lees JA, Harris SR, Tonkin-Hill G, et al. Fast and flexible bacterial genomic
531 epidemiology with PopPUNK. *Genome Research* 2019; **29**(2): 304-16.

532 23. Petit RA, 3rd, Read TD. Bactopia: a Flexible Pipeline for Complete Analysis of Bacterial
533 Genomes. *mSystems* 2020; **5**(4).
534 24. Zhang X, Payne M, Kaur S, Lan R. Improved Genomic Identification, Clustering, and
535 Serotyping of Shiga Toxin-Producing Escherichia coli Using Cluster/Serotype-Specific Gene
536 Markers. *Front Cell Infect Microbiol* 2021; **11**: 772574.
537 25. Strachan NJ, Rotariu O, Lopes B, et al. Whole Genome Sequencing demonstrates that
538 Geographic Variation of Escherichia coli O157 Genotypes Dominates Host Association. *Sci Rep*
539 2015; **5**: 14145.
540 26. Tarr GAM, Stokowski T, Shringi S, et al. Contribution and Interaction of Shiga Toxin
541 Genes to Escherichia coli O157:H7 Virulence. *Toxins* 2019; **11**(10).
542 27. Muller NF, Rasmussen DA, Stadler T. The Structured Coalescent and Its
543 Approximations. *Mol Biol Evol* 2017; **34**(11): 2970-81.
544 28. Dallman TJ, Ashton PM, Byrne L, et al. Applying phylogenomics to understand the
545 emergence of Shiga-toxin-producing Escherichia coli O157:H7 strains causing severe human
546 disease in the UK. *Microb Genom* 2015; **1**(3): e000029.
547 29. Byrne L, Dallman TJ, Adams N, Mikhail AFW, McCarthy N, Jenkins C. Highly
548 Pathogenic Clone of Shiga Toxin-Producing Escherichia coli O157:H7, England and Wales.
549 *Emerg Infect Dis* 2018; **24**(12): 2303-8.
550 30. Barth SA, Menge C, Eichhorn I, et al. The Accessory Genome of Shiga Toxin-Producing
551 Escherichia coli Defines a Persistent Colonization Type in Cattle. *Appl Environ Microbiol* 2016;
552 **82**(17): 5455-64.

553 31. Barth SA, Weber M, Schaufler K, Berens C, Geue L, Menge C. Metabolic Traits of
554 Bovine Shiga Toxin-Producing Escherichia Coli (STEC) Strains with Different Colonization
555 Properties. *Toxins (Basel)* 2020; **12**(6).
556 32. Tarr GAM, Rounds J, Vachon MS, Smith K, Medus C, Hedberg CW. Differences in risk
557 factors for transmission among Shiga toxin-producing Escherichia coli serogroups and stx
558 profiles. *J Infect* 2023; **87**(6): 498-505.
559
560

561 **SUPPLEMENTAL METHODS**

562 *STEC CASE DEFINITION*

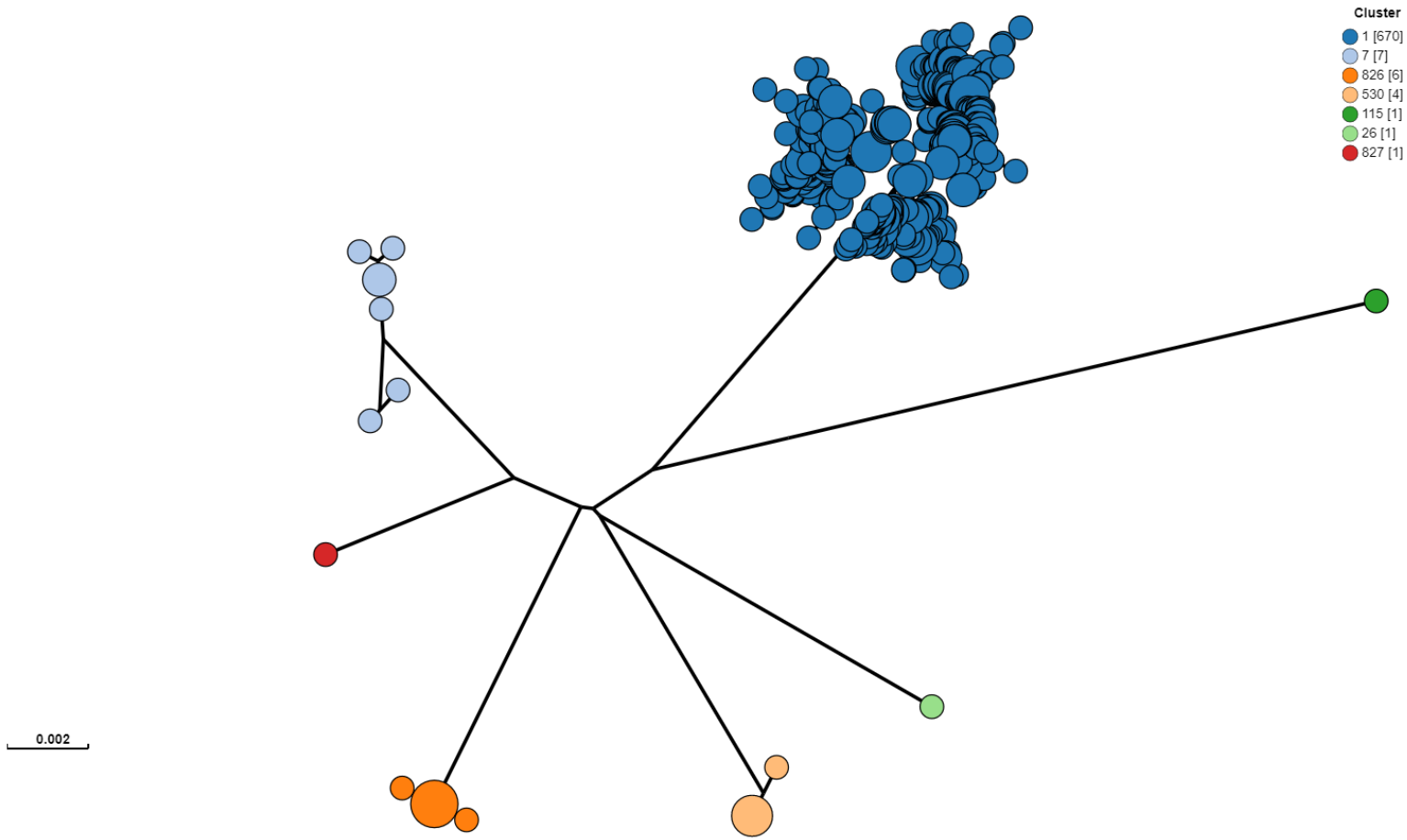
563 Alberta Health defines a confirmed case of Shiga toxin-producing *E. coli* (STEC), including *E.*
564 *coli* O157:H7, as STEC isolation or Shiga toxin antigen or nucleic acid detection. Clinical
565 illness, which may include diarrhea, bloody diarrhea, abdominal cramps, hemolytic uremic
566 syndrome, thrombocytopenia purpura, or pulmonary edema, may or may not be present.[33]

567

568 *SAMPLING FROM THE PRIOR DISTRIBUTION*

569 Results in our Bayesian phylodynamic analyses are drawn from posterior distributions, which are
570 influenced by both the data and the prior information we have about the system (Supplemental
571 Table S1). In order to confirm that our primary results were not overly influenced by our prior
572 assumptions, we conducted an analysis in which the sampling draws were made from the prior
573 distribution, as opposed to the posterior distribution. We graphed these results against the
574 sampling draws made from the posterior distributions from the four runs conducted for our
575 primary analysis (each performed with a different random seed). The comparison shows that the
576 draws from prior distribution differ markedly from the draws from the posterior distributions for
577 the model's key parameters (Supplemental Figure S2). From this, we concluded that our prior
578 assumptions were not overly influencing the results of the primary analysis.

579 **SUPPLEMENTAL FIGURES**



580

581 **Supplemental Figure S1. Genomic clustering of 690 Alberta *E. coli* O157 isolates.** Clustering performed from raw reads using

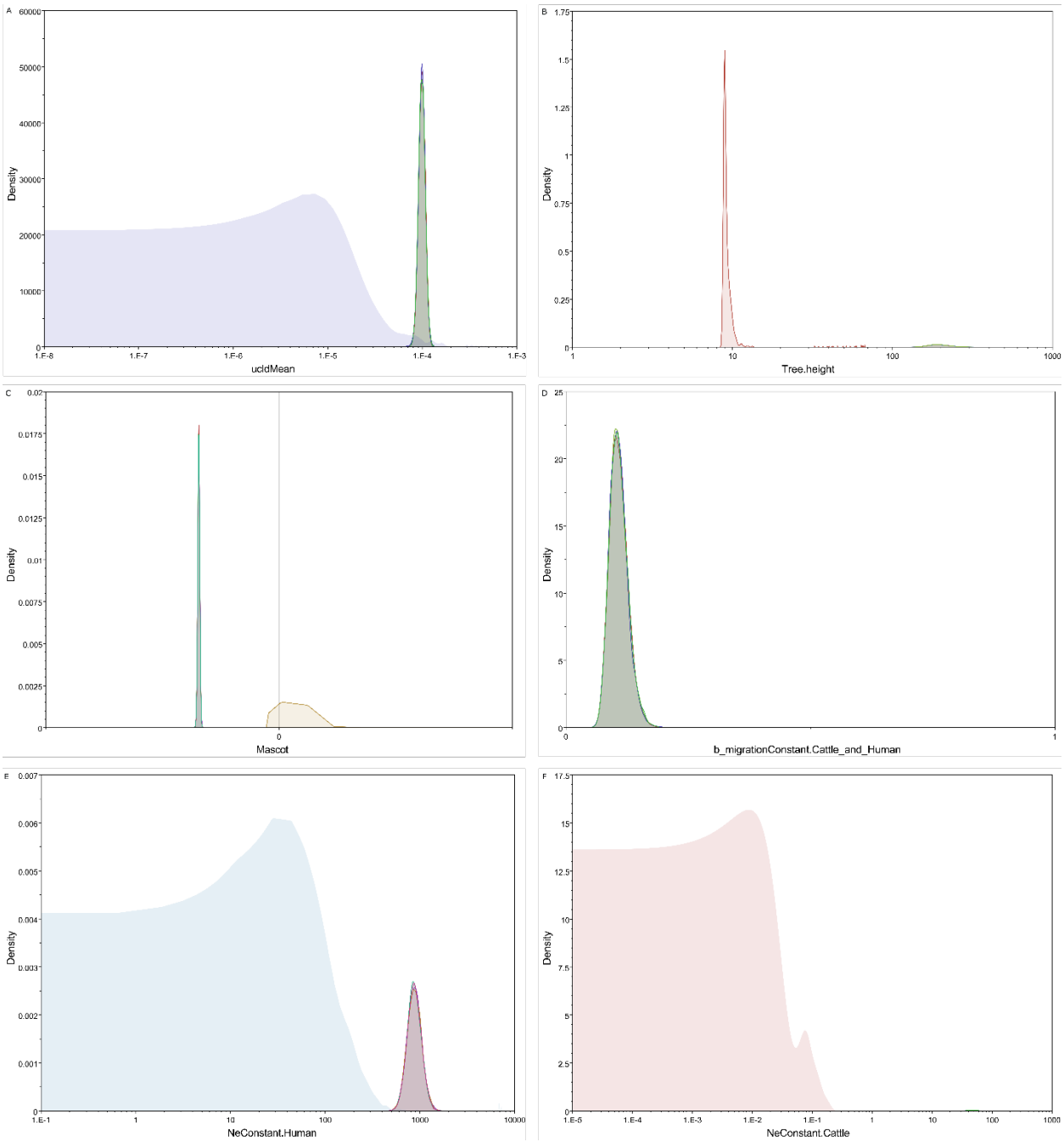
582 PopPUNK v2.5.0 with 10,146 *E. coli* reference genomes.[22] From the 246 isolates selected for sequencing for the study and 445

583 additional Alberta Health isolates included for contextualization, one isolate was removed prior to clustering analysis, because it was

584 identified through metadata review as an environmental (non-human, non-cattle) isolate. Cluster 1 included the Sakai and EDL933

585 reference strains. Clusters 826 and 827 were novel clusters. Isolates outside of Cluster 1 were excluded from all subsequent analyses.

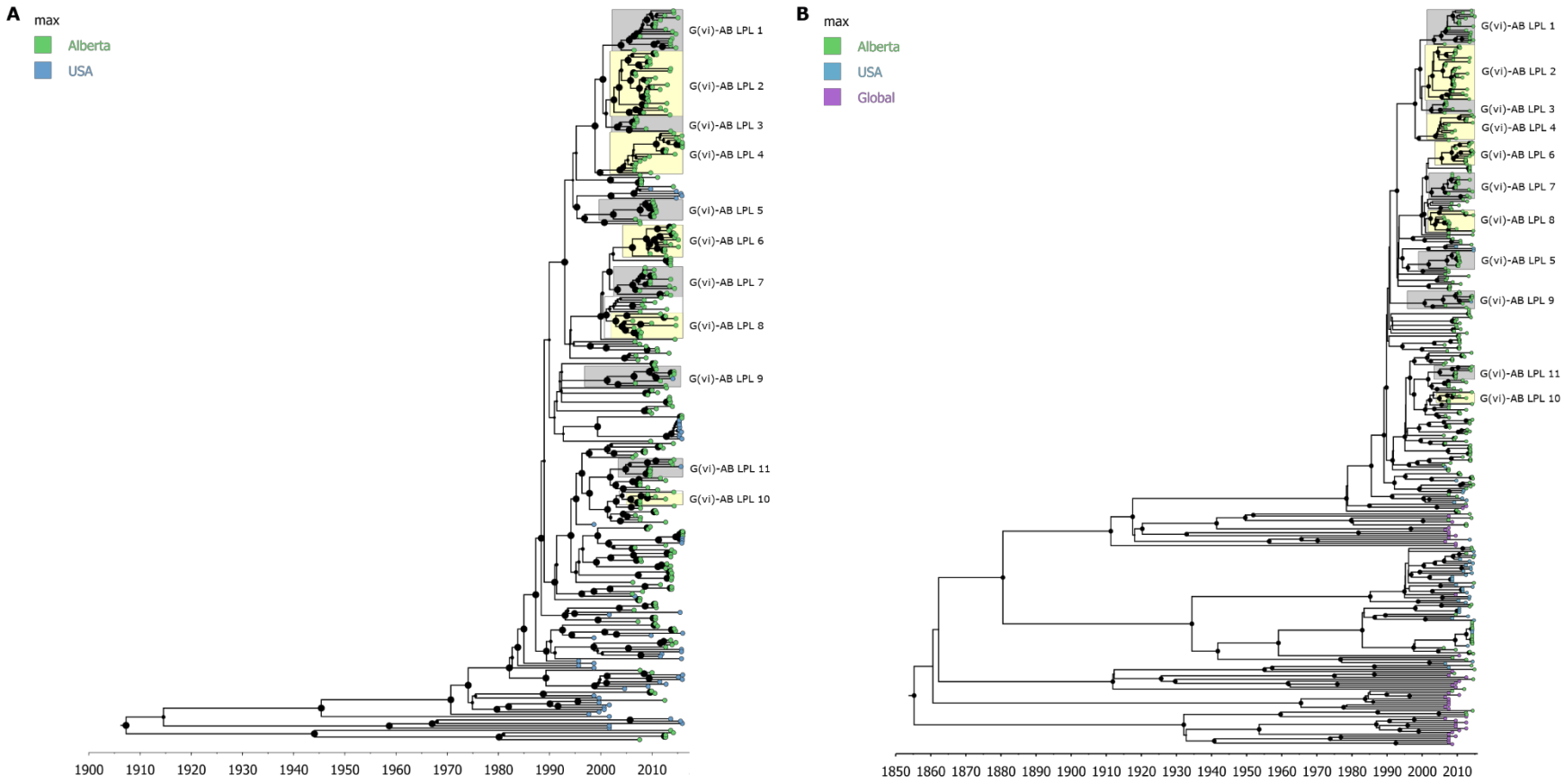
586



Supplemental Figure S2. Comparison of key parameters drawn from the posterior distributions of four runs using different starting seeds to draws from the prior distribution. The prior distributions for the clock rate (A, purple-grey), tree height (B, red), mascot (C, yellow), backward migration rate (D, not visible), human effective population size (E,

592 grey), and cattle effective population size (F, red) all differed substantially from the posterior

593 distributions of the four runs.



Supplemental Figure S3. Comparison of Alberta, U.S., and global *E. coli* O157:H7 isolates. Tips are colored based on isolate origin. Internal nodes are sized based on Bayesian posterior probability. LPLs are shaded and labeled as in Figure 3. Alberta isolates 2007-2016 and U.S. isolates 1996-2016 were analyzed for clade G only (A). Alberta, U.S., and other global isolates from 2007-2015 were analyzed for clades C through G (B). Two U.S. isolates, from 2014 and 2015, arose from Alberta LPLs 9 and 11, respectively. No global isolates were associated with Alberta LPLs.

600

601 ## SUPPLEMENTAL TABLES

602 **Supplemental Table S1.** Analyses conducted and model priors.

Analysis	Isolates Included	Tree Model	Substitution Model	Clock Model
Primary analysis [a]	Alberta 2007-2015 (n=229)	Structured coalescent with 2 demes; N_e initial value 10 and Weibull distribution (shape=1, scale=100); migration initial value 1.0 and Uniform distribution (minimum=0 and maximum=100)	HKY with empirical frequencies and discrete Gamma site model with 4 categories	Relaxed log-normal with initial value 3.5×10^{-5} and Log-Normal distribution ($M=3.5 \times 10^{-5}$, S=1.5, with mean in real space)
Alberta long-term persistence	Alberta 2007-2019 (n=459)[b]	Coalescent constant population with N_e initial value 10 and Weibull distribution (shape=1, scale=100)	HKY with empirical frequencies and discrete Gamma site model with 4 categories	Relaxed log-normal with initial value 1.5×10^{-4} and Log-Normal distribution ($M=1.5 \times 10^{-4}$, S=1.5, with mean in real space)
Clade G Alberta-U.S. importation	Alberta 2007-2016 (n=260)[b] U.S. 1996-2016 (n=56)	Structured coalescent with 2 demes; N_e initial value 10 and Weibull distribution (shape=1, scale=100); migration initial value 1.0 and Uniform distribution (minimum=0 and maximum=100)	HKY with empirical frequencies and discrete Gamma site model with 4 categories	Relaxed log-normal with initial value 1.5×10^{-4} and Log-Normal distribution ($M=1.5 \times 10^{-4}$, S=1.5, with mean in real space)
Global circulation [c]	Alberta 2007-2015 (n=281)[b] U.S. 2007-2015 (n=46) Global 2007-2015 (n=47)	Structured coalescent with 3 demes; N_e initial value 10 and Weibull distribution (shape=1, scale=100); migration initial value 1.0 and Uniform distribution (minimum=0 and maximum=100)	HKY with empirical frequencies and discrete Gamma site model with 4 categories	Relaxed log-normal with initial value 1.5×10^{-4} and Log-Normal distribution ($M=1.5 \times 10^{-4}$, S=1.5, with mean in real space)

603 [a] Primary analysis was run using four different random seeds to confirm that all converged to the same solution. The four runs were combined to produce the final
604 maximum clade credibility tree and state transition estimates. Model priors from this analysis were also used in an analysis in which draws were taken from the
605 prior distribution, as opposed to the posterior distribution, to confirm that the final results were not overly influenced by the choice of priors.

606 [b] Outbreaks were down-sampled to avoid bias.

607 [c] Clades A and B were excluded from this analysis given the very small number of isolates available for them from any locale.

608

609 **Supplemental Table S2.** Estimated migrations from structured coalescent analysis of Alberta,

610 U.S., and global isolates, excluding clades A and B.

Migration Direction	Mean	95% HPD
Alberta to Alberta	478	357, 624
Alberta to Global	26	19, 34
Alberta to U.S.	66	50, 78
Global to Alberta	0	0, 2
Global to Global	39	24, 54
Global to U.S.	0	0, 0
U.S. to Alberta	87	8, 147
U.S. to Global	0	0, 1
U.S. to U.S.	49	0, 106

611 Abbreviation: HPD, highest posterior density interval

612

613 **SUPPLEMENTAL REFERENCES**

614 33. Alberta Health. Escherichia coli verotoxigenic infections. 2021.

615 https://open.alberta.ca/publications/escherichia-coli-verotoxigenic-infections (accessed January

616 25 2024).

617