Microbiology and Infectious Disease
Epidemiology and Global Health

Reviewed Preprint
v1 • August 29, 2024
Not revised

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

Gillian AM Tarr, Linda Chui, Kim Stanford, Emmanuel W Bumunang, Rahat Zaheer, Vincent Li, Stephen B Freedman, Chad R Laing, Tim A McAllister

Division of Environmental Health Sciences, School of Public Health, University of Minnesota, MMC807, Room 1240 Mayo, 420 Delaware Street SE, Minneapolis, MN, 55455, USA • Alberta Precision Laboratories, Alberta Public Health, Room 1B2.19 Walter Mackenzie Health Sciences Centre, 8440-112th Street, Edmonton, Alberta, T6G 2J2, Canada • Department of Laboratory Medicine and Pathology, University of Alberta, Edmonton, Alberta, Canada • Department of Biological Sciences, University of Lethbridge, 4401 University Drive, Lethbridge, Alberta, T1K 3M4, Canada • Agriculture and Agri-Food Canada, Lethbridge Research and Development Centre, 5403 – 1 Avenue South. P.O. Box 3000, Lethbridge, Alberta, T1J 4B1, Canada • 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 • National Center for Animal Diseases Lethbridge Laboratory, Canadian Food Inspection Agency, 225090 Township Road 91, Lethbridge County, Alberta, T1J 5R7, Canada

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*.

Conclusions

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

eLife assessment

This work revealed numerous distinct lineages that evolved within a local human population in Alberta, Canada, leading to persistent cases of E. coli O157:H7 infections for over a decade and highlighting the ongoing involvement of local cattle in disease transmission, as well as the possibility of intermediate hosts and environmental reservoirs. This is a **useful** study that also showed a shift towards more virulent stx2a-only strains becoming predominant in the local lineages. The paper's evidence supporting the role played by cattle in the transmission system of E. coli O157:H7 in Alberta is currently **incomplete**, based on potential sampling issues associated with the selection of isolates as raised by the reviewers.

https://doi.org/10.7554/eLife.97643.1.sa3

Introduction

Several areas around the globe experience exceptionally high incidence of Shiga toxin-producing *Escherichia coli* (STEC), including the virulent serotype *E. coli* O157:H7. These include Scotland,[1] Ireland,[2] Argentina,[3] and the Canadian province of Alberta.[4] All are home to large populations of agricultural ruminants, STEC's primary reservoir. However, there are many regions with similar ruminant populations where STEC incidence is unremarkable. What differentiates high risk regions is unclear. Moreover, with systematic STEC surveillance only conducted in limited parts of the world,[5] there may be unidentified regions with exceptionally high disease burden.

STEC infections can arise from local reservoirs, transmitted through food, water, direct animal contact, or contact with contaminated environmental matrices. The most common reservoirs include domesticated ruminants such as cattle, sheep, and goats. While STEC has been isolated from a variety of other animal species and outbreaks have been linked to species such as deer[6] and swine,[7] it is unclear what roles they play as maintenance or intermediate hosts. STEC infections can be imported through food items traded nationally and internationally, as has been seen with *E. coli* O157:H7 outbreaks in romaine lettuce from the United States.[8] Secondary transmission is believed to cause approximately 15% of cases, but the pathogen is not believed to be sustained long-term through person-to-person transmission alone.[9,10]

The mix of STEC infection sources in a region directly influences public health measures needed to control disease burden. Living near cattle and other domesticated ruminants has been linked to STEC incidence, particularly for *E. coli* O157:H7.[2,11–15] These studies suggest an important role for local reservoirs in STEC epidemiology. A comprehensive understanding of STEC's disease ecology would enable more effective investigations into potential local transmission systems and ultimately their control. Here, we take a phylodynamic, genomic epidemiology approach to more precisely discern the role of the cattle reservoir in the dynamics of *E. coli* O157:H7 human infections. We focus on the high incidence region of Alberta, Canada to provide insight into characteristics that make the pathogen particularly prominent in such regions.

Methods

Study Design and Population

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

In addition to the 246 isolates for the primary analysis, we contextualized our findings with three additional sets of *E. coli* O157:H7 isolates (**Figure 1**): 445 from Alberta Health from 2009 to 2019 and already sequenced as part of other public health activities; 152 from the U.S. from 1999 to 2016; and 54 from elsewhere around the world between 2007 and 2015. The additional Alberta Health isolates were sequenced by the National Microbiology Laboratory (NML)-Public Health Agency of Canada (Winnipeg, Manitoba, Canada) as part of PulseNet Canada activities. Isolates sequenced by the NML for 2018 and 2019 constituted the majority of reported *E. coli* O157:H7 cases for those years (217 of 247; 87.9%). U.S. isolates were considered separately from other global isolates, as the U.S. is Alberta's most frequent international trade partner, with both processed beef and live cattle crossing the border. U.S. isolates from 1999 to 2009 and global isolates were identified from previous literature,[20] and U.S. isolates from 2010 to 2016 were randomly selected from *E. coli* O157:H7 sequences available through the U.S. CDC's PulseNet BioProject PRJNA218110.

This study was approved by the University of Calgary Conjoint Health Research Ethics Board, #REB19-0510. A waiver of consent was granted, and all case data were deidentified.

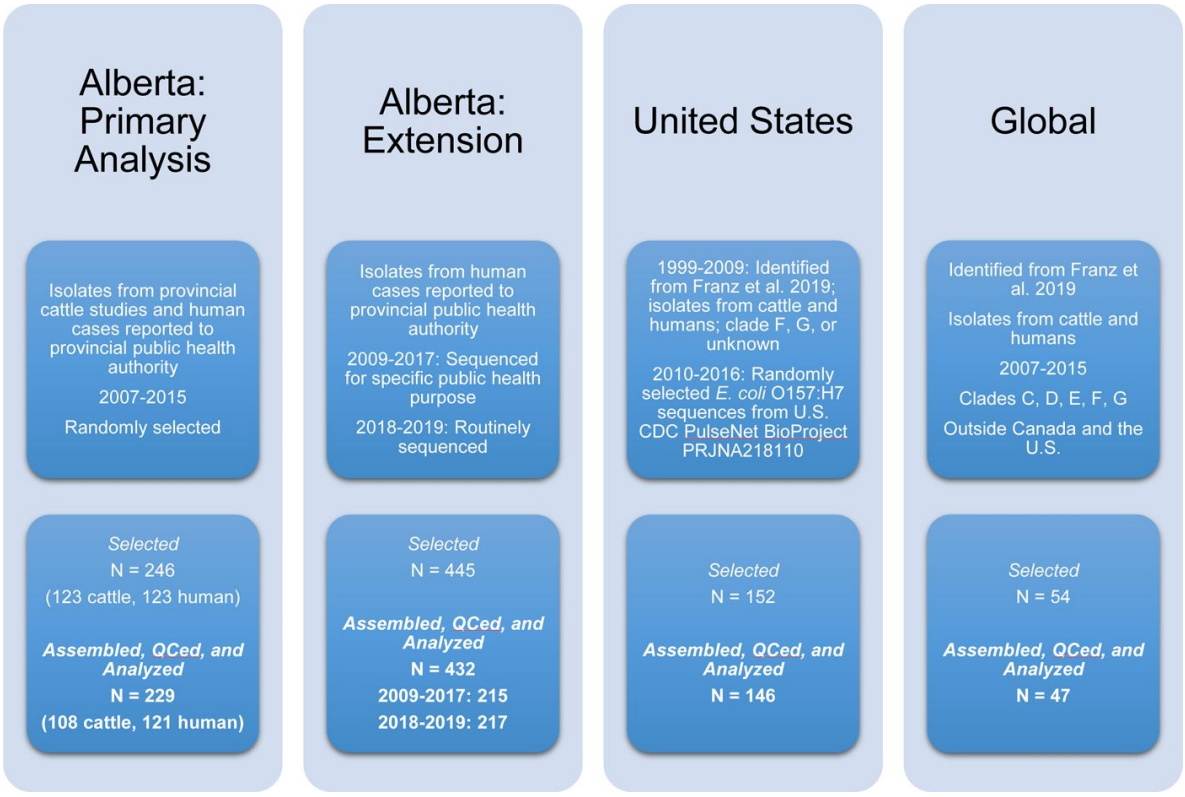


Figure 1.

***E. coli* O157:H7 isolates selected for the study and included in analysis.**

Four sets of isolates, all originating from cattle or humans, were included in the study.

Whole Genome Sequencing, Assembly, and Initial Phylogeny

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

PopPUNK v2.5.0 was used to cluster Alberta isolates and identify any outside the O157:H7 genomic cluster (**Supplemental Figure S1**).[22] For assembling and quality checking (QC) all sequences, we used the Bactopia v3.0.0 pipeline.[23] This pipeline performed an initial QC step on the reads using FastQC v0.12.1, which trimmed adapters and read ends with quality lower than 6 and discarded reads after trimming with overall quality scores lower than 10. None of the isolates were eliminated during this step for low read quality. We used the Shovill v1.1.0 assembler within the Bactopia pipeline to *de novo* assemble the Unicycler contigs for the primary analysis and raw reads from the supplementary datasets. Bactopia generated a quality report on the assemblies, which we assessed based on number of contigs (<500), genome size (≥5.1 Mb), N50 (>30,000), and L50 (≤50). Low-quality assemblies were removed. This included 1 U.S. sequence, for which 2 FASTQ files had been attached to a single BioSample identifier; the other sequence for the isolate passed all quality checks and remained in the analysis. Additionally, 16 sequences from the primary analysis dataset and 4 from the extended Alberta data had a total length <5.1 Mb. These sequences corresponded exactly to those identified by the PopPUNK analysis to be outside the primary *E. coli* O157:H7 genomic cluster. Finally, although all isolates were believed to be of cattle or clinical origin during initial selection, detailed metadata review identified 1 isolate of environmental origin in the primary analysis dataset and 8 that had been isolated from food items in the extended Alberta data. These were excluded. We used STECFinder v1.1.0[24] to determine Shiga toxin gene (*stx*) profile and confirm the *E. coli* O157:H7 serotype using the *wzy* or *wzx* O157 O-antigen genes and detection of the H7 H-antigen. After processing, we had 229 isolates (121 human, 108 cattle) in our primary sample, 432 additional Alberta Health isolates, 146 U.S. isolates, and 47 global isolates (**Figure 1**, Supplemental Data File).

Bactopia's Snippy workflow, which incorporates Snippy v4.6.0, Gubbins v3.3.0, and IQTree v2.2.2.7, followed by SNP-Sites v2.5.1, were used to generate a core genome SNP alignment with recombinant blocks removed. The maximum likelihood phylogeny of the core genome SNP alignment generated by IQTree was visualized in Microreact v251. The number of core SNPs between isolates was calculated using PairSNP v0.3.1. Clade was determined based on the presence of at least one defining SNP for the clade as published previously.[25] Isolates were identified to the subclade level [e.g. G(vi)] when both clade and subclade SNPs were present and the clade level (e.g. G) when only clade SNPs were present.

Phylodynamic and Statistical Analyses

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

Local persistent lineages (LPLs) were identified based on following criteria: 1) a single lineage of the MCC tree with a most recent common ancestor (MRCA) with ≥95% posterior probability; 2) all isolates <30 core SNPs from one another; 3) contained at least 1 cattle isolate; 4) contained ≥5 isolates; and 5) the isolates were collected at sampling events (for cattle) or reported (for humans) over a period of at least 1 year. From non-LPL isolates, we estimated the number of local transient isolates vs. imported isolates. For the 121 human *E. coli* O157:H7 isolates in the primary sample, we determined what portion belonged to local persistent lineages (LPL) and what portion were likely to be from local transient *E. coli* O157:H7 populations vs. imported. Human isolates within the LPLs were enumerated (n = 44). The 77 human isolates outside LPLs included 58 clade G(vi) isolates and 19 non-G(vi) isolates. Based on the MCC tree from the primary analysis, none of the non-G(vi) human isolates was likely to have been closely related to an isolate from the Alberta cattle population, suggesting that all 19 were imported. As a proportion of all non-LPL human isolates, these 19 constituted 24.7%. While it may be possible that all clade G(vi) isolates were part of a local evolving lineage, it is also possible that exchange of both cattle and food from other locations was causing the regular importation of clade G(vi) strains and infections. Thus, we used the proportion of non-LPL human isolates outside the G(vi) clade to estimate the proportion of non-LPL human isolates within the G(vi) clade that were imported; i.e., 58 × 24.7% = 14. We then conducted a similar exercise for cattle isolates.

To contextualize our results in terms of ongoing human disease burden, we created a timed phylogeny using a constant, unstructured coalescent model of the 229 Alberta isolates from the primary analysis and the additional Alberta Health isolates. Outbreaks were down-sampled to avoid biasing the tree by randomly selecting 1 to 2 isolates per outbreak; as such, only 230 of the 432 additional isolates were included in the analysis (**Supplemental Table S1**). We identified LPLs as above, and leveraged the near-complete sequencing of isolates from 2018 and 2019 to calculate the proportion of reported human cases associated with LPLs.

We created a timed phylogeny of Alberta isolates and U.S. isolates from 1996 to 2016 to test whether the LPLs or Alberta's dominant *E. coli* O157:H7 clade (G) were linked to U.S. ancestors (**Supplemental Table S1**). We also created a timed phylogeny of temporally overlapping Alberta, U.S., and global isolates from 2007 to 2015, excluding clades A and B, which were too limited to 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.

Results

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

The Majority of Clinical Cases Evolved from Local Cattle Lineages

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

Figure 2.

Maximum likelihood core SNP tree of the 854 *E. coli* O157:H7 isolates referenced in the study.

This includes 229 randomly sampled cattle and human isolates from Alberta, Canada, from 2007 through 2015; 432 additional Alberta isolates sequenced as part of public health activities from 2009 through 2019; and 152 isolates from the U.S. and 47 isolates from elsewhere around the globe from 1996 to 2016 to examine international transmission history. Clade is shown in the coloration of the tips on the 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 ring. The tree was rooted at clade A. Clade G constituted the majority of isolates.

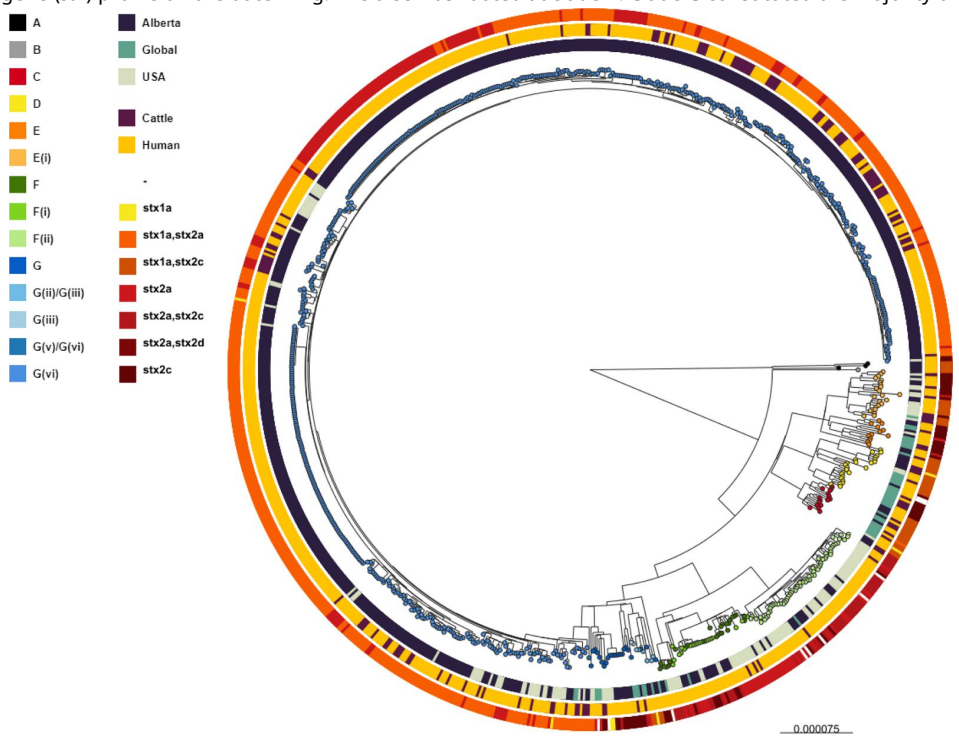


Table 1.

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**

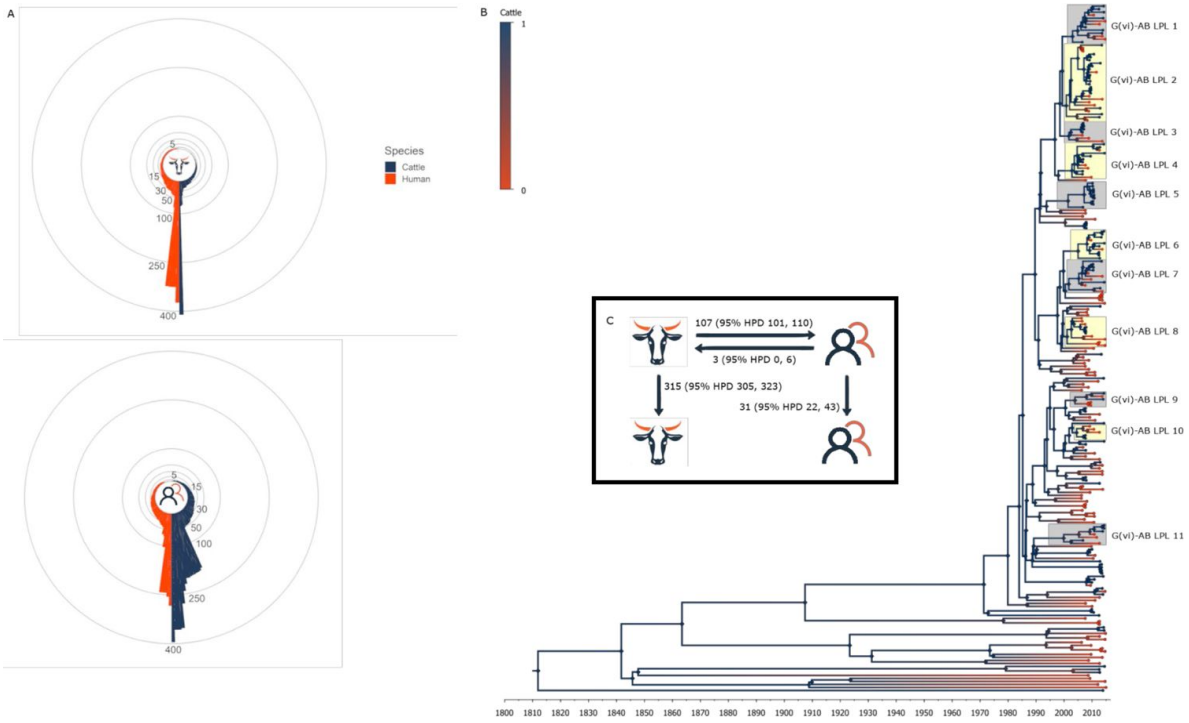


Figure 3.

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

The phylogeny generated by our primary structured coalescent analysis indicated cattle were the primary reservoir, with a high probability that the hosts at nodes along the tree's backbone were cattle (**Figure 3b**). The root was estimated at 1812 (95% HPD 1748, 1870). The most recent common ancestor (MRCA) of clade G(vi) strains in Alberta was inferred to be a cattle strain, dated to 1971 (95% HPD 1961, 1980). With our assumption of a relaxed molecular clock, the mean clock rate for the core genome was estimated at 1.00×10^{-4} (95% HPD 8.45×10^{-5}, 1.18×10^{-4}) substitutions/site/year. The effective population size, N_e, of the human *E. coli* O157:H7 population was estimated as 913 (95% HPD 620, 1232), and for cattle as 49 (95% HPD 32, 67). We estimated 107 (95% HPD 101, 111) human lineages arose from cattle lineages, and 31 (95% HPD 22, 43) arose from other human lineages (**Figure 3c**). In other words, 77.5% of human lineages arose from cattle lineages. We observed minimal influence of our choice of priors (**Supplemental Figure S2**, Supplemental Text).

Local Persistent Lineages Account for the Majority of Ongoing Human Disease

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

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

To understand long-term persistence, we expanded the phylogeny with additional Alberta Health isolates from 2009 to 2019 (**Supplemental Table S1**). Six of the 11 LPLs identified in our primary analysis continued to cause disease during the 2016 to 2019 period (**Figure 4a**). With most of the cases reported during 2018 and 2019 sequenced, we were able to estimate the proportion of reported *E. coli* O157:H7 associated with LPLs. Of 217 sequenced cases reported during these two years, 162 (74.7%; 95% CI 68.3%, 80.3%) arose from Alberta's LPLs. The *stx* profile of LPL isolates shifted as compared to the primary analysis, with 83 (51.2%) of the LPL isolates encoding only *stx2a* and the rest *stx1a/stx2a* (**Figure 4b**). Among the 55 non-LPL isolates during 2018-2019, the *stx2c*-only profile emerged with 16 (29.1%) isolates, and *stx2a*-only was found in only 6 (10.9%) cases.

All 5 large (≥10 cases) sequenced outbreaks in Alberta during the study period were within clade G(vi). LPLs gave rise to 3 large outbreaks, accounting for 117 cases, including 83 from an extended outbreak by a single strain, defined as isolates within 5 SNPs of one another, during 2018 and 2019 (**Figure 4a**). The two large outbreaks that did not arise from LPLs both occurred in 2014 and were responsible for 164 cases.

International Importation Does Not Explain Alberta's Current Disease Burden

Only 2 U.S. isolates coincided with Alberta LPLs, specifically G(vi)-AB LPL 9 in 2014 and G(vi)-AB LPL 11 in 2015 (**Supplemental Figure S3**). Isolates in these two LPLs from Alberta dated to 2007 and 2009, respectively, and were identified multiple times up to and including during the 2018-2019 period (**Figure 4a**). There was no evidence of early U.S. ancestors of LPLs. No LPL

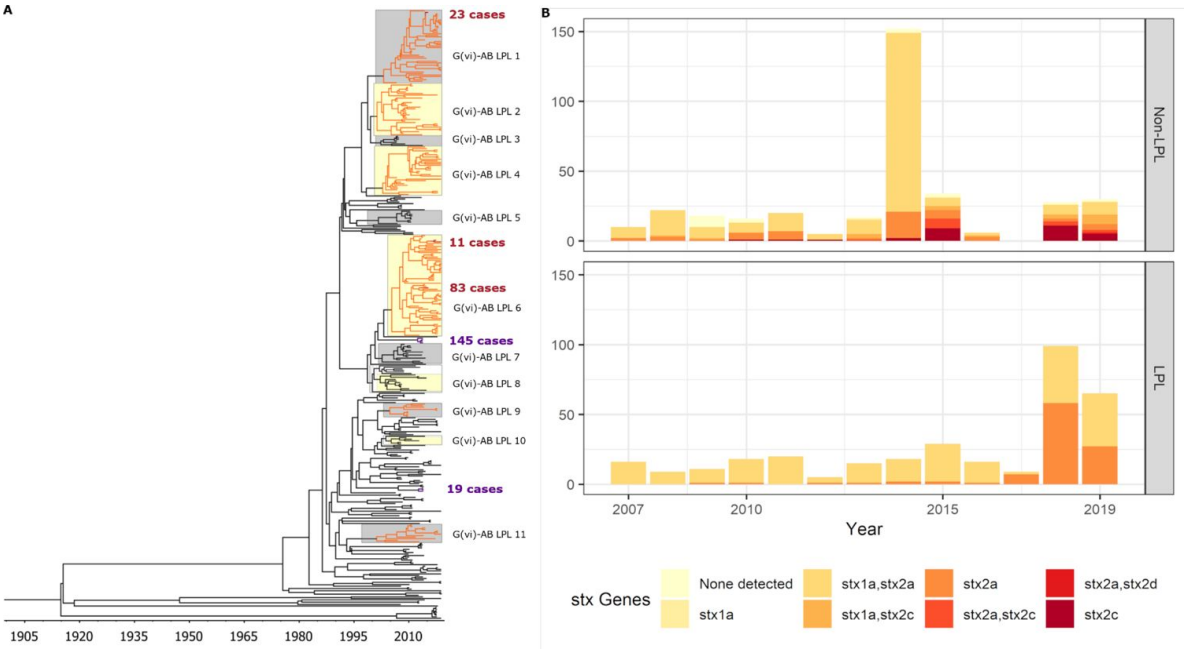


Figure 4.

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 represented associated with each outbreak are shown in red (LPL-associated outbreaks) or purple (non-LPL-associated outbreaks) text. 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 period (2007-2015) to the later study period (2016-2019), with more of the virulent *stx2a*-only profile observed in 2018 and 2019 than in previous years (B). The peak in sequences in 2014 is due to two outbreaks; routine sequencing began in 2018 and 2019, accounting for the rise in sequenced cases during those years.

contained a global isolate. Based on migration events calculated from the tree, we estimated that 15.4% of combined human and cattle Alberta lineages were imported (**Supplemental Table S2**). Sequences from outside North America were separated from Alberta sequences by a median of 325 (IQR 288-349) SNPs. Including U.S. and global isolates in the phylogeny did not alter the LPLs identified, though some minor rearrangement of the tree was observed (**Supplemental Figure S3**).

Discussion

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

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

The cattle-human transitions we estimated were based on structured coalescent theory,[27] which we used throughout our analyses. This approach is similar to phylogeographic methods that have previously been applied to *E. coli* O157:H7.[20] We inferred the full backbone of the Alberta *E. coli* O157:H7 phylogeny as arising from cattle, consistent with the postulated global spread of the pathogen via ruminants.[20] Our estimate of the origin of the serotype, at 1812 (95% HPD 1748, 1870), was somewhat earlier than previous estimates, but consistent with global (1890; 95% HPD 1845, 1925)[20] and United Kingdom (1840; 95% HPD 1817, 1855)[28] studies that used comparable methods. Our dating of Alberta's G(vi) clade to 1971 (95% HPD 1961, 1980) also corresponds to proposed migrations of clade G into Canada from the U.S. in 1965-1977.[20] Our study thus adds to the growing body of work on the larger history of *E. coli* O157:H7, providing an in-depth examination of the G(vi) clade.

Our identification of the 11 local persistent lineages (LPLs) is significant in demonstrating that the majority of Alberta's reported *E. coli* O157:H7 illnesses are of local origin. Our definition ensured that every LPL had an Alberta cattle strain and at least 5 isolates separated by >1 year, making the importation of the isolates in a lineage highly unlikely. Further supporting the evolution of the LPLs within Alberta, all 11 LPLs were in clade G(vi), several were phylogenetically related with MRCAs dating to the late 1990s, and few non-Alberta isolates fell within LPLs. The two U.S. isolates associated with Alberta LPLs may reflect Alberta cattle that were slaughtered in the U.S. Thus, we are confident that the identified LPLs represent locally evolving lineages and potential persistent sources of disease.

Based on our LPL analysis, we estimated only 27% of human and 10% of cattle *E. coli* O157:H7 isolates were imported. This was consistent with the overall importation estimate of 15% for all Alberta lineages from our global structured coalescent analysis. While these estimates may appear low given the recent focus on row crops and other produce as potential vehicles of disease,[8] 26% of sporadic STEC infections have been attributed to animal contact and the farm environment, with a further 19% to pink or raw meat.[10] Similarly, 24% of *E. coli* O157 outbreaks in the U.S. were attributed to beef, animal contact, water, or other environmental sources.[9] In Alberta, these are all inherently local exposures, given that 90% of beef consumed in Alberta is produced and/or processed there. Even person-to-person transmission, responsible for 15% of sporadic cases and 16% of outbreaks,[9,10] includes secondary transmission from cases infected from local sources, which may explain our estimate of 23% for person-to-person transmission. To our knowledge, our study provides the first comprehensive determination of local vs. imported status for *E. coli* O157:H7 cases. Similar studies in regions of both high and moderate incidence would provide further insight into the role of localization on *E. coli* O157:H7 incidence.

Of the 11 lineages we identified as LPLs during the 2007-2015 period, 6 were also identified in the 2016-2019 period. During the initial period, 36% of human cases were linked to an LPL, and 6.1% carried only *stx2a*. The risk of HUS increases in strains of STEC carrying only *stx2a*, relative to *stx1a/stx2a*,[26] meaning the earlier LPL population had few of the high-virulence strains. In 2018 and 2019, the 6 long-term LPLs were associated with both greater incidence and greater virulence, encompassing 75% of human cases with more than half of LPL isolates carrying only *stx2a*. The cause of this shift remains unclear, though shifts toward greater virulence in *E. coli* O157:H7 populations have been seen elsewhere.[29] The growth and diversity of G(vi)-AB LPL 1 and G(vi)-AB LPL 6 in the later period suggest these lineages were in stable reservoirs or adapted easily to new niches. Identifying these reservoirs could yield substantial insights into disease prevention, given the significant portion of illnesses caused by persistent strains.

We developed a novel measure of persistence for use in this study, specifically for the purposes of identifying lineages that pose an ongoing threat to public health in a specific region.Persistence has been defined variably in the literature, for example as shedding of the same strain for at least 4 months.[30] Most recently, the U.S. CDC has identified the first Recurring, Emergent, and Persistent (REP) STEC strain, REPEXH01, which has been detected since 2017 in over 600 cases. REPEXH01strains are within 21 allele differences of one another (*https://www.cdc.gov/ncezid/dfwed/outbreak-response/rep-strains/repexh01.html*). Given that we used high resolution SNP analysis rather than MLST, we used a difference of <30 SNPs to define persistent lineages. Supporting the persistence we have observed, the REPEXH01 strain is also an *E. coli* O157:H7 strain; however, O157:H7 was defined as sporadic in a German study using the 4-month shedding definition, which may be due to ecological differences.[30] Understanding microbial drivers of persistence is an active field of research, with early findings suggesting a correlation of STEC persistence to the accessory genome and traits such as biofilm formation and nutrient metabolism.[30,31] Our approach to studying persistence was specifically designed for longitudinal sampling in high-incidence regions and may be useful for others attempting to identify sources that disproportionately contribute to disease burden.

Our analysis was limited to only cattle and humans. However, small ruminants (e.g., sheep, goats) have also been identified as important STEC reservoirs,[12,15,25] and Alberta has experienced outbreaks linked to swine.[7] Had isolates from a wider range of potential reservoirs been available, we would have been able to elucidate more clearly the roles that various hosts and common sources of infection play in local transmission. This may help explain the 3 human-to-cattle predicted transmissions, which could be erroneous. We also limited our analysis only to *E. coli* O157:H7 despite the growing importance of non-O157 STEC as historical multi-species collections of non-O157 isolates are lacking. As serogroups differ meaningfully in exposures,[32] our results may not be generalizable beyond the O157 serogroup. Finally, we were not able to

estimate the impact of strain migration between Alberta and the rest of Canada, because metadata for publicly-available *E. coli* O157:H7 sequences from Canada was limited, such that we could not be sure they were from outside Alberta.

E. coli O157:H7 infections are a pressing public health problem in many high incidence regions around the world including Alberta, where a recent childcare outbreak caused >300 illnesses. In the majority of sporadic cases, and even many outbreaks,[9] the source of infection is unknown, making it critical to understand the disease ecology of *E. coli* O157:H7 at a system level. Here we have identified a high proportion of human cases arising from cattle lineages and a low proportion of imported cases. Local transmission systems, including intermediate hosts and environmental reservoirs, need to be elucidated to develop management strategies that reduce the risk of STEC infection. In Alberta, local transmission is dominated by a single clade, and over the extended study period, persistent lineages caused an increasing proportion of disease. The local lineages with long-term persistence are of particular concern because of their increasing virulence, yet they also present opportunity as larger, more stable targets for reservoir identification and control.

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.

Supplemental Methods

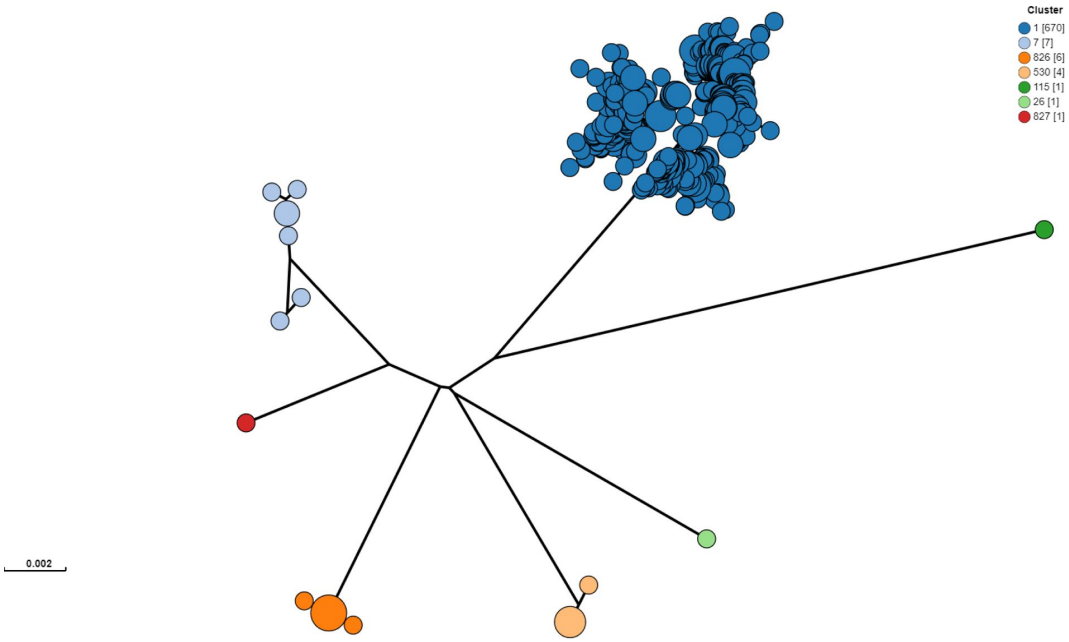
STEC Case Definition

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

Sampling from the Prior Distribution

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

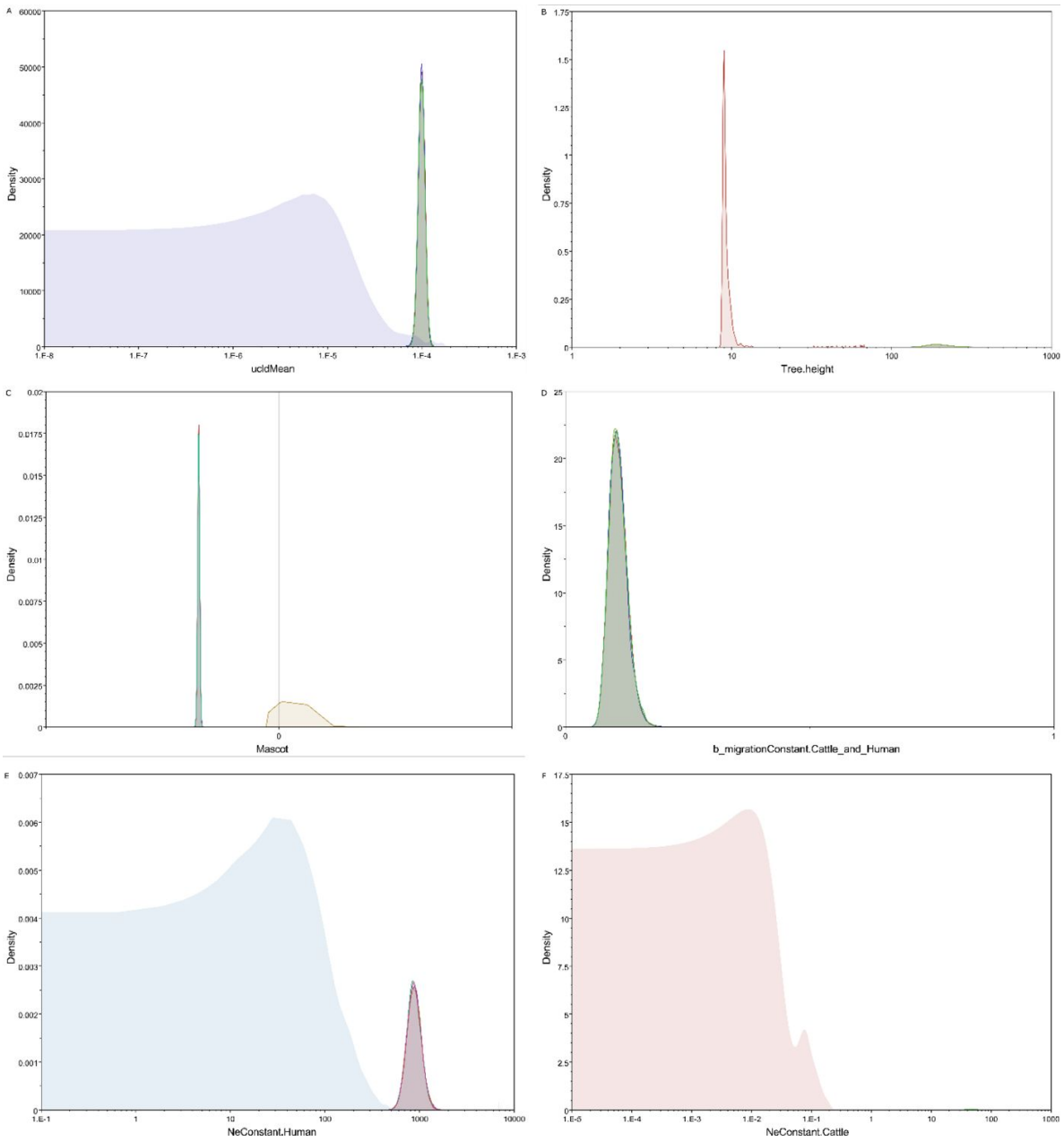
Supplemental Figures



Supplemental Figure S1.

Genomic clustering of 690 Alberta *E. coli* O157 isolates.

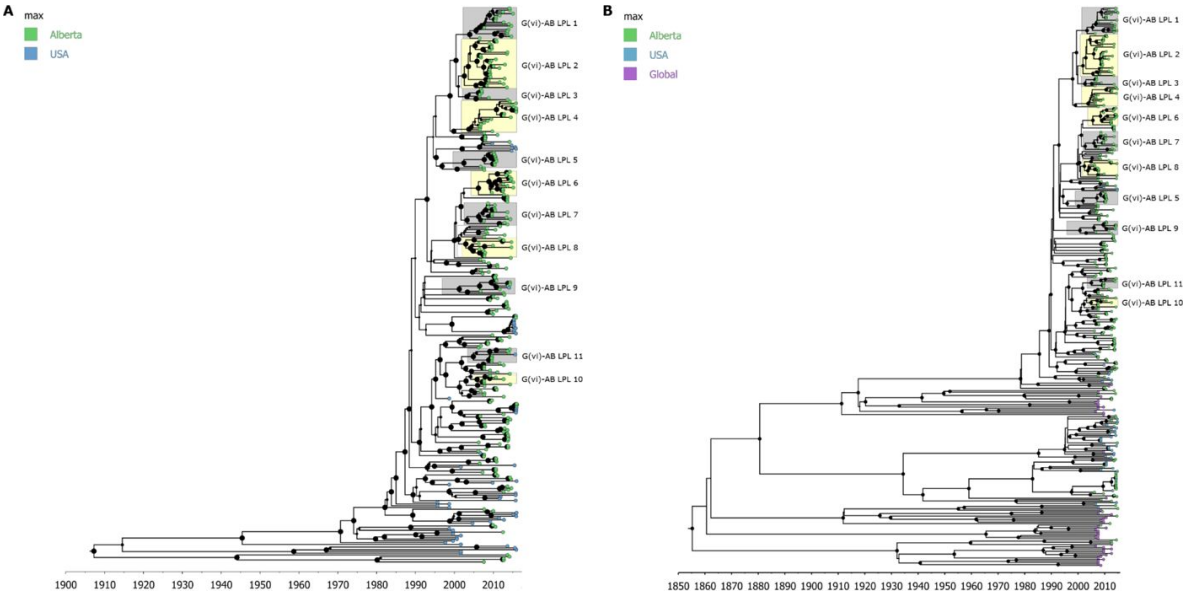
Clustering performed from raw reads using PopPUNK v2.5.0 with 10,146 *E. coli* reference genomes.[22] From the 246 isolates selected for sequencing for the study and 445 additional Alberta Health isolates included for contextualization, one isolate was removed prior to clustering analysis, because it was identified through metadata review as an environmental (non-human, non-cattle) isolate. Cluster 1 included the Sakai and EDL933 reference strains. Clusters 826 and 827 were novel clusters. Isolates outside of Cluster 1 were excluded from all subsequent analyses.



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

Supplemental Tables

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 $3.5x10^{-5}$ and Log-Normal distribution ($M=3.5x10^{-5}$, 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 $1.5x10^{-4}$ and Log-Normal distribution ($M=1.5x10^{-4}$, 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 $1.5x10^{-4}$ and Log-Normal distribution ($M=1.5x10^{-4}$, 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 $1.5x10^{-4}$ and Log-Normal distribution ($M=1.5x10^{-4}$, S=1.5, with mean in real space)

[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 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 prior distribution, as opposed to the posterior distribution, to confirm that the final results were not overly influenced by the choice of priors.

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

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

Supplemental Table S1.

Analyses conducted and model priors.

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

Abbreviation: HPD, highest posterior density interval

Supplemental Table S2.

Estimated migrations from structured coalescent analysis of Alberta, U.S., and global isolates, excluding clades A and B.

References

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

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

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

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

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

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

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

8. Bottichio L, Keaton A, Thomas D, et al. (2020) **Shiga Toxin-Producing Escherichia coli Infections Associated With Romaine Lettuce-United States, 2018** *Clin Infect Dis* **71**:e323–e30

9. Tack DM, Kisselburgh HM, Richardson LC, et al. (2021) **Shiga Toxin-Producing Escherichia coli Outbreaks in the United States, 2010-2017** *Microorganisms* **9**

10. Kintz E, Brainard J, Hooper L, Hunter P (2017) **Transmission pathways for sporadic Shiga-toxin producing E. coli infections: A systematic review and meta-analysis** *Int J Hyg Environ Health* **220**:57–67

11. Jaros P, Cookson AL, Campbell DM, et al. (2013) **A prospective case-control and molecular epidemiological study of human cases of Shiga toxin-producing Escherichia coli in New Zealand** *BMC Infect Dis* **13**

12. Elson R, Grace K, Vivancos R, et al. (2018) **A spatial and temporal analysis of risk factors associated with sporadic Shiga toxin-producing Escherichia coli O157 infection in England between 2009 and 2015** *Epidemiol Infect* **146**:1928–39

13. Frank C, Kapfhammer S, Werber D, Stark K, Held L (2008) **Cattle density and Shiga toxin-producing Escherichia coli infection in Germany: increased risk for most but not all serogroups** *Vector Borne Zoonotic Dis* **8**:635–43

14. Friesema IH, Van De Kassteele J CM DEJ, Heuvelink AE, Van Pelt W (2011) **Geographical association between livestock density and human Shiga toxin-producing Escherichia coli O157 infections** *Epidemiol Infect* **139**:1081–7

15. Mulder AC, van de Kassteele J, Heederik D, Pijnacker R, Mughini-Gras L, Franz E (2020) **Spatial Effects of Livestock Farming on Human Infections With Shiga Toxin-Producing Escherichia coli O157 in Small but Densely Populated Regions: The Case of the Netherlands** *Geohealth* **4**

16. Stephens TP, McAllister TA, Stanford K (2009) **Perineal swabs reveal effect of super shedders on the transmission of Escherichia coli O157:H7 in commercial feedlots** *J Anim Sci* **87**:4151–60

17. Stanford K, Hannon S, Booker CW, Jim GK (2014) **Variable efficacy of a vaccine and direct-fed microbial for controlling Escherichia coli O157:H7 in feces and on hides of feedlot cattle** *Foodborne Pathog Dis* **11**:379–87

18. Stanford K, Gibb D, McAllister TA (2013) **Evaluation of a shelf-stable direct-fed microbial for control of Escherichia coli O157 in commercial feedlot cattle** *Canadian Journal of Animal Science* **93**:535–42

19. Stanford K, Johnson RP, Alexander TW, McAllister TA, Reuter T (2016) **Influence of Season and Feedlot Location on Prevalence and Virulence Factors of Seven Serogroups of Escherichia coli in Feces of Western-Canadian Slaughter Cattle** *PLoS One* **11**

20. Franz E, Rotariu O, Lopes BS, et al. (2019) **Phylogeographic Analysis Reveals Multiple International transmission Events Have Driven the Global Emergence of Escherichia coli O157:H7** *Clin Infect Dis* **69**:428–37

21. Bumunang EW, Zaheer R, Stanford K, et al. (2022) **Genomic Analysis of Shiga Toxin-Producing E. coli O157 Cattle and Clinical Isolates from Alberta, Canada** *Toxins (Basel)* **14**

22. Lees JA, Harris SR, Tonkin-Hill G, et al. (2019) **Fast and flexible bacterial genomic epidemiology with PopPUNK** *Genome Research* **29**:304–16

23. Petit RA, Read TD (2020) **Bactopia: a Flexible Pipeline for Complete Analysis of Bacterial Genomes** *mSystems* **5**

24. Zhang X, Payne M, Kaur S, Lan R (2021) **Improved Genomic Identification, Clustering, and Serotyping of Shiga Toxin-Producing Escherichia coli Using Cluster/Serotype-Specific Gene Markers** *Front Cell Infect Microbiol* **11**

25. Strachan NJ, Rotariu O, Lopes B, et al. (2015) **Whole Genome Sequencing demonstrates that Geographic Variation of Escherichia coli O157 Genotypes Dominates Host Association** *Sci Rep* **5**

26. Tarr GAM, Stokowski T, Shringi S, et al. (2019) **Contribution and Interaction of Shiga Toxin Genes to Escherichia coli O157:H7 Virulence** *Toxins* **11**

27. Muller NF, Rasmussen DA, Stadler T (2017) **The Structured Coalescent and Its Approximations** *Mol Biol Evol* **34**:2970–81

28. Dallman TJ, Ashton PM, Byrne L, et al. (2015) **Applying phylogenomics to understand the emergence of Shiga-toxin-producing Escherichia coli O157:H7 strains causing severe human disease in the UK** *Microb Genom* **1**

29. Byrne L, Dallman TJ, Adams N, Mikhail AFW, McCarthy N, Jenkins C (2018) **Highly Pathogenic Clone of Shiga Toxin-Producing Escherichia coli O157:H7, England and Wales** *Emerg Infect Dis* **24**:2303–8

30. Barth SA, Menge C, Eichhorn I, et al. (2016) **The Accessory Genome of Shiga Toxin-Producing Escherichia coli Defines a Persistent Colonization Type in Cattle** *Appl Environ Microbiol* **82**:5455–64

31. Barth SA, Weber M, Schaufler K, Berens C, Geue L, Menge C (2020) **Metabolic Traits of Bovine Shiga Toxin-Producing Escherichia Coli (STEC) Strains with Different Colonization Properties** *Toxins (Basel)* **12**

32. Tarr GAM, Rounds J, Vachon MS, Smith K, Medus C, Hedberg CW (2023) **Differences in risk factors for transmission among Shiga toxin-producing Escherichia coli serogroups and stx profiles** *J Infect* **87**:498–505

33. Alberta Health (2021) **Alberta Health. Escherichia coli verotoxigenic infections. 2021. https://open.alberta.ca/publications/escherichia-coli-verotoxigenic-infections (accessed January 25 2024).**

33. Alberta Health (2021) **Alberta Health. Escherichia coli verotoxigenic infections. 2021. https://open.alberta.ca/publications/escherichia-coli-verotoxigenic-infections (accessed January 25 2024).**

Editors

Reviewer #1 (Public Review):

Summary:

A nice study trying to identify the relationship between E. coli O157 from cattle and humans in Alberta, Canada.

Strengths:

(1) The combined human and animal sampling is a great foundation for this kind of study.

(2) Phylogenetic analyses seem to have been carried out in a high-quality fashion.

Weaknesses:

I think there may be a problem with the selection of the isolates for the primary analysis. This is what I'm thinking:

(1) Transmission analyses are strongly influenced by the sampling frame.

(2) While the authors have randomly selected from their isolate collections, which is fine, the collections themselves are not random.

(3) The animal isolates are likely to represent a broad swathe of diversity, because of the structured sampling of animal reservoirs undertaken (as I understand it).

(4) The human isolates are all from clinical cases. Clinical cases of the disease are likely to be closely related to other clinical cases, because of outbreaks (either detected, or undetected), and the high ascertainment rate for serious infections.

(5) Therefore, taking an equivalent number of animal and clinical isolates, will underestimate the total diversity in the clinical isolates because the sampling of the clinical isolates is less "independent" (in the statistical sense) than sampling from the animal isolates.

(6) This could lead to over-estimating of transmission from cattle to humans.

(7) "We hypothesize that the large proportion of disease associated with local transmission systems is a principal cause of Alberta's high E. coli O157:H7 incidence" - this seems a bit tautological. There is a lot of O157 because there's a lot of transmission. What part of the fact it is local means that it is a principal cause of high incidence? It seems that they've observed a high rate of local transmission, but the reasons for this are not apparent, and hence the cause of Alberta's incidence is not apparent. Would a better conclusion not be that "X% of STEC in Alberta is the result of transmission of local variants"? And then, this poses a question for future epi studies of what the transmission pathway is.

https://doi.org/10.7554/eLife.97643.1.sa2

Reviewer #2 (Public Review):

This study identified multiple locally evolving lineages transmitted between cattle and humans persistently associated with E. coli O157:H7 illnesses for up to 13 years. Furthermore, this study mentions a dramatic shift in the local persistent lineages toward strains with the more virulent stx2a-only profile. The authors hypothesized that this phenomenon is the large proportion of disease associated with local transmission systems is a principal cause of Alberta's high E. coli O157:H7 incidence. These opinions more effectively explain the role of the cattle reservoir in the dynamics of E. coli O157:H7 human infections.

(1) The authors acknowledge the possibility of intermediate hosts or environmental reservoirs playing a role in transmission. Further discussion on the potential roles of other animal species commonly found in Alberta (e.g., sheep, goats, swine) could enhance the understanding of the transmission dynamics. Were isolates from these species available for analysis? If not, the authors should clearly state this limitation.

(2) The focus on E. coli O157:H7 is understandable given its prominence in Alberta and the availability of historical data. However, a brief discussion on the potential applicability of the findings to non-O157 STEC serogroups, and the limitations therein, would be beneficial. Are there reasons to believe the transmission dynamics would be similar or different for other serogroups?

(3) The authors briefly mention the need for elucidating local transmission systems to inform management strategies. A more detailed discussion on specific public health interventions

that could be targeted at the identified LPLs and their potential reservoirs would strengthen the paper's impact.

(4) "Understanding the relationship between specific risk factors and E. coli O157:H7 infections is essential for developing effective prevention strategies. Have case-control or cohort studies been conducted to assess the correlation between identified risk factors and the incidence of E. coli O157:H7 infections? What methodologies were employed to control for potential confounders in these studies?"

(5) The study's findings are noteworthy, particularly in the context of E. coli O157:H7 epidemiology. However, the extent to which these results can be replicated across different temporal and geographical settings remains an open question. It would be constructive for the authors to provide additional data that demonstrate the replication of their sampling and sequencing experiments under varied conditions. This would address concerns regarding the specificity of the observed patterns to the initial study's parameters.

https://doi.org/10.7554/eLife.97643.1.sa1

Author response:

> ***Reviewer #1 (Public Review):***
>
> *Summary:*
>
> *A nice study trying to identify the relationship between E. coli O157 from cattle and humans in Alberta, Canada.*
>
> *Strengths:*
>
> *(1) The combined human and animal sampling is a great foundation for this kind of study.*
>
> *(2) Phylogenetic analyses seem to have been carried out in a high-quality fashion.*
>
> *Weaknesses:*
>
> *I think there may be a problem with the selection of the isolates for the primary analysis. This is what I'm thinking:*
>
> *(1) Transmission analyses are strongly influenced by the sampling frame.*
>
> *(2) While the authors have randomly selected from their isolate collections, which is fine, the collections themselves are not random.*
>
> *(3) The animal isolates are likely to represent a broad swathe of diversity, because of the structured sampling of animal reservoirs undertaken (as I understand it).*
>
> *(4) The human isolates are all from clinical cases. Clinical cases of the disease are likely to be closely related to other clinical cases, because of outbreaks (either detected, or undetected), and the high ascertainment rate for serious infections.*
>
> *(5) Therefore, taking an equivalent number of animal and clinical isolates, will underestimate the total diversity in the clinical isolates because the sampling of the clinical isolates is less "independent" (in the statistical sense) than sampling from the animal isolates.*
>
> *(6) This could lead to over-estimating of transmission from cattle to humans.*

We appreciate the reviewer's careful thoughts about our sampling strategy. We agree with points (1) and (2), and we will provide additional details on the animal collections as requested.

We agree with point (3) in theory but not in fact. As shown in Figure 3a, the cattle isolates were very closely related, despite the temporal and geographic breadth of sampling within Alberta. The median SNP distance between cattle sequences was 45 (IQR 36-56), compared to 54 (IQR 43-229) SNPs between human sequences from cases in Alberta during the same years. Additionally, as shown in Figure 2, only clade A and B isolates – clades that diverge substantially from the rest of the tree – were dominated by human cases in Alberta. We will better highlight this evidence in the revision.

We agree with the reviewer in point (4) that outbreaks can be an important confounder of phylogenetic inference. This is why we down-sampled outbreaks (based on genetic relatedness, not external designation) in our extended analyses (lines 192-194). We did not do this in the primary analysis, because there were no large clusters of identical isolates. Figure 3b shows a limited number of small clusters; however, clustered cattle isolates outnumbered clustered human isolates, suggesting that any bias would be in the opposite direction the reviewer suggests. Regarding severe cases being oversampled among the clinical isolates, this is absolutely true and a limitation of all studies utilizing public health reporting data. We will make this limitation to generalizability clearer in the discussion. However, as noted above, clinical isolates were more variable than cattle isolates, so it does not appear to have heavily biased the analysis.

We disagree with the reviewer on point (5). While the bias toward severe cases could make the human isolates less independent, the relative sampling proportions are likely to induce greater distance between clinical isolates than cattle isolates, which is exactly what we observe (see response to point (3) above). Cattle are E. coli O157:H7's primary reservoir, and humans are incidental hosts not able to sustain infection chains long-term. Not only is the bacteria prevalent among cattle, cattle are also highly prevalent in Alberta. Thus, even with 89 sampling points, we are still capturing a small proportion of the E. coli O157:H7 in the province. Being able to sample only a small proportion of cattle's E. coli O157:H7 increases the likelihood of only sampling from the center of the distribution, making extreme cases such as that shown at the very bottom of the tree in Figure 3b, rare and important. In comparison, sampling from human cases constitutes a higher proportion of human infections relative to cattle, and is therefore more representative of the underlying distribution, including extremes. We will add this point to the limitations. As with the clustering above, if anything, this outcome would have biased the study away from identifying cattle as the primary reservoir. Additionally, the relatively small proportion of cattle sampled makes our finding that 15.7% of clinical isolates were within 5 SNPs of a cattle isolate, the distance most commonly used to indicate transmission for E. coli O157:H7, all the more remarkable.

Because of the aforementioned points, we disagree with the reviewer's conclusion in point (6). We believe transmission from cattle-to-humans is likely underestimated for the reasons given above. Not only do all prior studies indicate ruminants as the primary reservoirs of E. coli O157:H7, and humans as only incidental hosts, our specific data do not support the reviewer's individual contentions. That said, we will conduct a sensitivity analysis as recommended to determine the impact of sampling and inclusion of the small clusters on our primary findings.

> *(7) We hypothesize that the large proportion of disease associated with local transmission systems is a principal cause of Alberta's high E. coli O157:H7 incidence" - this seems a bit tautological. There is a lot of O157 because there's a lot of transmission. What part of the fact it is local means that it is a principal cause of high incidence? It seems that they've observed a high rate of local transmission, but the reasons for this are*

> *not apparent, and hence the cause of Alberta's incidence is not apparent. Would a better conclusion not be that "X% of STEC in Alberta is the result of transmission of local variants"? And then, this poses a question for future epi studies of what the transmission pathway is.*

The reviewer is correct, and the suggestion for the direction of future studies was our intent with this statement. We will revise it.

> ***Reviewer #2 (Public Review):***
>
> *This study identified multiple locally evolving lineages transmitted between cattle and humans persistently associated with E. coli O157:H7 illnesses for up to 13 years. Furthermore, this study mentions a dramatic shift in the local persistent lineages toward strains with the more virulent stx2a-only profile. The authors hypothesized that this phenomenon is the large proportion of disease associated with local transmission systems is a principal cause of Alberta's high E. coli O157:H7 incidence. These opinions more effectively explain the role of the cattle reservoir in the dynamics of E. coli O157:H7 human infections.*
>
> *(1) The authors acknowledge the possibility of intermediate hosts or environmental reservoirs playing a role in transmission. Further discussion on the potential roles of other animal species commonly found in Alberta (e.g., sheep, goats, swine) could enhance the understanding of the transmission dynamics. Were isolates from these species available for analysis? If not, the authors should clearly state this limitation.*

We will expand the discussion of other species in Alberta, as suggested, including other livestock, wildlife, and the potential role of birds and flies. Unfortunately, we did not have sequences available from other species, and we will add this to the limitations. Sequences from other species may be available from sequences collected by others, which as we note in the limitations do not have sufficient metadata to assign them to Alberta vs. the rest of Canada. While we have requested this data, we have been unsuccessful in obtaining it. We will continue to pursue it.

> *(2) The focus on E. coli O157:H7 is understandable given its prominence in Alberta and the availability of historical data. However, a brief discussion on the potential applicability of the findings to non-O157 STEC serogroups, and the limitations therein, would be beneficial. Are there reasons to believe the transmission dynamics would be similar or different for other serogroups?*

We appreciate this comment and will expand our discussion of relevance to non-O157 STEC. Other authors have proposed that transmission dynamics differ, and studies of STEC risk factors, including our own, support this. However, there has been very little direct study of non-O157 transmission dynamics and there is even less cross-species genomic and metadata available for non-O157 isolates of concern.

> *(3) The authors briefly mention the need for elucidating local transmission systems to inform management strategies. A more detailed discussion on specific public health interventions that could be targeted at the identified LPLs and their potential reservoirs would strengthen the paper's impact.*

We agree with the reviewer that this would be a good addition to the manuscript. The public health implications for control are several and extend to non-STEC reportable zoonotic enteric infections, such as Campylobacter and Salmonella. We will add a discussion of these.

> *(4) Understanding the relationship between specific risk factors and E. coli O157:H7 infections is essential for developing effective prevention strategies. Have case-control or cohort studies been conducted to assess the correlation between identified risk factors and the incidence of E. coli O157:H7 infections? What methodologies were employed to control for potential confounders in these studies?*

Yes, there have been several case-control studies of reported cases. Many of these are referenced in the discussion in terms of the contribution of different sources to infection. However, we will add a more explicit discussion of risk factors.

> *(5) The study's findings are noteworthy, particularly in the context of E. coli O157:H7 epidemiology. However, the extent to which these results can be replicated across different temporal and geographical settings remains an open question. It would be constructive for the authors to provide additional data that demonstrate the replication of their sampling and sequencing experiments under varied conditions. This would address concerns regarding the specificity of the observed patterns to the initial study's parameters.*

We appreciate the reviewer's comment, as we are currently building on this analysis with an American dataset with different types of data available than were used in this study. We will add a discussion of this. We will also be adding a sensitivity analysis to the manuscript simulating a different sampling approach, which should also be informative to this question.

https://doi.org/10.7554/eLife.97643.1.sa0