Assessing the genomic relatedness and evolutionary rates of persistent verotoxigenic Escherichia coli serotypes within a closed beef herd in Canada

Lu Ya Ruth Wang1, Cassandra C. Jokinen2, Chad R. Laing3, Roger P. Johnson4, Kim Ziebell4 and Victor P. J. Gannon1,*

Abstract
Verotoxigenic Escherichia coli (VTEC) are food- and water-borne pathogens associated with both sporadic illness and outbreaks of enteric disease. While it is known that cattle are reservoirs of VTEC, little is known about the genomic variation of VTEC in cattle, and whether the variation in genomes reported for human outbreak strains is consistent with individual animal or group/herd sources of infection. A previous study of VTEC prevalence identified serotypes carried persistently by three consecutive cohorts of heifers within a closed herd of cattle. This present study aimed to: (i) determine whether the genomic relatedness of bovine isolates is similar to that reported for human strains associated with single source outbreaks, (ii) estimate the rates of genome change among dominant serotypes over time within a cattle herd, and (iii) identify genomic features of serotypes associated with persistence in cattle. Illumina MiSeq genome sequencing and genotyping based on allelic and single nucleotide variations were completed, while genome change over time was measured using Bayesian evolutionary analysis sampling trees. The accessory genome, including the non-protein-encoding intergenic regions (IGRs), virulence factors, antimicrobial-resistance genes and plasmid gene content of representative persistent and sporadic cattle strains were compared using Fisher’s exact test corrected for multiple comparisons. Herd strains from serotypes O6:H34 (n=22), O22:H8 (n=30), O108:H8 (n=39), O139:H19 (n=44) and O157:H7 (n=106) were readily distinguishable from epidemiologically unrelated strains of the same serotype using a similarity threshold of 10 or fewer allele differences between adjacent nodes. Temporal-cohort clustering within each serotype was supported by date randomization analysis. Substitutions per site per year were consistent with previously reported values for E. coli; however, there was low branch support for these values. Acquisition of the phage-encoded Shiga toxin 2 gene in serotype O22:H8 was observed. Pan-genome analyses identified accessory regions that were more prevalent in persistent serotypes (P<0.05) than in sporadic serotypes. These results suggest that VTEC serotypes from a specific cattle population are highly clonal with a similar level of relatedness as human single-source outbreak-associated strains, but changes in the genome occur gradually over time. Additionally, elements in the accessory genomes may provide a selective advantage for persistence of VTEC within cattle herds.

DATA SUMMARY
FASTQ sequences have been deposited in the National Center for Biotechnology Information Sequence Read Archive under BioProject number PRJNA565946. Eight supplementary figures and three supplementary tables are available with the online version of this article.
INTRODUCTION

Verotoxigenic Escherichia coli (VTEC) are food- and water-borne pathogens associated with sporadic cases and outbreaks of severe enteric disease, of which non-O157 serotypes are recognized as emerging and previously under-surveyed pathogens [1]. With the continual integration of whole-genome sequencing (WGS) into public-health laboratories and national surveillance programs [2–6], the degree of genomic variation among VTEC isolates is now commonly used for cluster identification during outbreak investigations [7–9], routine surveillance [10–12], and retrospective characterization of isolates to address broader questions of epidemiology and population structure [13–15].

The current literature supports a range of ‘clonality thresholds’ for the clustering of E. coli [10, 11, 15–18], with a provisional benchmark of ≤10 SNPs or allele differences between epidemiologically related strains [19]. Public Health England initiates outbreak investigations for clusters comprising at least five isolates within 5 SNPs genetic distance and a 30 day isolation time frame [20]. For VTEC O157:H7, PulseNet Canada has observed ≤5 single nucleotide variants (SNVs) or <10 whole-genome multilocus sequence typing (wgMLST) allele differences between outbreak-related isolates [8].

While cattle are known to be important reservoirs of VTEC responsible for human illness [21, 22], there is a paucity of estimates for the genomic variation of VTEC isolated from cattle and whether it is congruent with strain variation observed for human clinical strains from common source outbreaks. Similarly, the evolutionary rate of the organism within cattle is poorly understood. A study of genome-scale rates of evolutionary change in 16 species of bacteria, excluding E. coli, reported rates from $10^{-8}$ to $10^{-5}$ substitutions per site per year [23]. Highly clonal bacteria, such as Mycobacterium tuberculosis, have reported rates of 4 SNPs per 4 years [24], while more divergent bacteria such as Helicobacter pylori can have more than 30 SNPs per year [25]. Within E. coli, the global evolutionary rate of enterotoxigenic E. coli was $3.7 \times 10^{-7}$ to $1.1 \times 10^{-6}$ substitutions per site per year or 2 to 5.5 SNPs per genome per year [26]. Estimates for VTEC O157:H7 of human and bovine origin and the genetically similar Shigella were 2.6 and 3.6 SNPs per genome per year ($8.5 \times 10^{-7}$ substitutions per site per year), respectively [15, 27]. Stoesser et al. [28] estimated an evolutionary rate of $2.46 \times 10^{-7}$ mutations per site per year (95 % confidence interval (CI) 2.18$\times10^{-7}$ to 2.75$\times10^{-7}$) or 1 mutation per genome per year (95 % CI 0.89 to 1.12) for the extra-intestinal and globally distributed E. coli clone ST131. As rates are generally estimated based on public genomes from spatially and temporally divergent populations, we examined whether similar rates of change are observed in populations that are spatially and temporally restricted, as would be the case during common source outbreaks.

Persistence of enteric bacterial pathogens, here defined as the long-term detection of strains of high molecular similarity within an environment, may contribute to repeated food or environment contamination and foodborne outbreaks [13, 29, 30]. This attribute has been exploited by targeted mitigation efforts, such as the seek-and-destroy process employed for Listeria monocytogenes [31]. Persistence of serotypes and clones in distinct cattle populations has been previously reported in O157:H7 and non-O157 [32–34], attributed to bacteria–host interactions such as enhanced colonization [35], and defined by a shared repertoire of virulence-associated genes in the accessory genome [36]. However, accessory content outside of conventional virulence genes may also have clinical associations, as reported in Campylobacter [37] and Clostridium difficile [38].

In a previous study of VTEC prevalence within a closed beef herd, we identified several serotypes that were persistent within and across multiple cohort-years [39]. In this study, we aimed to examine the relatedness of persistent VTEC using this highly homogenous population as a baseline. Genomic relatedness within each persistent serotype was assessed using: wgMLST, pan-genome and reference-derived SNP analysis; accessory genome analysis including virulence, antibiotic-resistance gene and plasmid content; and Bayesian inference of time-scaled phylogenies and evolutionary rates. The objectives were to: (i) determine whether the genomic relatedness of bovine isolates is similar to that reported for human strains associated with single source outbreaks, (ii) estimate the rates of genome change among dominant serotypes over

Impact Statement

Cattle are recognized as the most important reservoir of verotoxigenic Escherichia coli (VTEC). In this study, genomic changes among five VTEC serotypes (O6:H34, O22:H8, O139:H19, O108:H8 and O157:H7) characterized in a cattle herd over a 3 year period were analysed. Overall, strains from the same animal, cohort and year were the most highly related, sharing a similar level of genomic variation as that reported for single-source human outbreak strains. Within-serotype clustering by cohort-year was noted, as was continuous evolution of clusters in time-scaled phylogenies. While the degree of genomic relatedness within serotypes changed slowly, distinct clusters of strains were seen to emerge and dominate or disappear over time. Based on these results, the threshold used for inferring genomic relatedness among strains should be considerably greater for long-term pathogen surveillance studies than outbreak investigations. Furthermore, the acquisition/loss of a Shiga toxin 2 phage by members of different strain clusters within the O22:H8 population was observed with the Stx2-positive clade dominating in the last year. Lastly, by comparing the genomes of persistent serotypes and sporadically isolated serotypes, we identified genomic features that are associated with herd persistence. These attributes are potential targets for use in the control of VTEC in cattle.
Table 1. Number of herd strains analysed in this study

| Serotype | No. of herd strains | Year            | n   |
|----------|---------------------|-----------------|-----|
| O6:H34   | 22                  | 2012–2013       | 13  |
|          |                     | 2013–2014       | 3   |
|          |                     | 2014–2015       | 6   |
| O22:H8   | 30                  | 2012–2013       | 9   |
|          |                     | 2013–2014       | 6   |
|          |                     | 2014–2015       | 15  |
| O108:H8  | 39                  | 2012–2013       | 14  |
|          |                     | 2013–2014       | 9   |
|          |                     | 2014–2015       | 16  |
| O139:H19 | 44                  | 2012–2013       | 18  |
|          |                     | 2013–2014       | 11  |
|          |                     | 2014–2015       | 15  |
| O157:H7  | 106                 | 1995            | 4   |
|          |                     | 1996            | 45  |
|          |                     | 1997            | 46  |
|          |                     | 1998            | 11  |

Wang et al., Microbial Genomics 2020;6

Table 1. Number of herd strains analysed in this study
2012–2013 = cohort-year 1; 2013–2014 = cohort-year 2; 2014–2015 = cohort-year 3 (Wang et al., 2018) [39].

Methods

Bacterial isolates

Herd characteristics and sample collection for VTEC have been described previously [39]. Briefly, 38 yearling heifers from three consecutive cohorts (10 to 16 individuals per cohort-year) from the Canadian Food Inspection Agency (CFIA)-National Centre for Animal Disease (NCAD) (Lethbridge, Alberta, Canada) Angus-Hereford cross research herd were sampled for VTEC on a monthly basis between April 2012 and March 2015. The closed herd shared a size and management system similar to those found in beef cow-calf herds in western Canada [34]. Four serotypes persistently isolated were selected for serotype-specific analysis in the present study – O6:H34, O22:H8, O108:H8 and O139:H38; an additional set of VTEC O157:H7 isolates collected from the same herd between 1995 and 1998 was also included for analysis (Table 1).

WGS

The MasterPure DNA purification kit (Epicentre) or the DNeasy blood and tissue kit (Qiagen) was used for extraction of genomic DNA from pure cultures. Nextera XT libraries were sequenced on a MiSeq system (Illumina) using 2×300 bp paired-end reads. A minimum raw sequence coverage of 40× based on an estimated genome size of 5 Mb was calculated using the Integrated Rapid Infectious Analysis (IRIDA) platform [40]. Sequence quality was assessed using BioNumerics v7.6.2 with the following parameters: AvgQuality ≥30, N50 ≥100,000, NrContigs <200, length 4.5 to 5.5 Mb, NrConsensus (congruent allele calls by assembly-based and assembly-free algorithms) >3800, NrDifferent 0 (alleles with no overlap between assembly-based and assembly-free algorithms), core per cent ≥90 %. Paired-end read data and whole-genome assemblies for outgroup strains used in this study were downloaded from the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) and GenBank, respectively. A complete list of strains and corresponding metadata is available in Table S1, available with the online version of this article. The Shoiviil pipeline v1.0.1 [41] was used for genome assembly using SPAdes v3.11 [42]. Other functions included genome size estimation, FASTQ subsampling, adaptor trimming, read correction, merging of paired-end reads and assembly correction using: MASH [43], seqtk [44], kmc 2 [45], Trimomatic [46], Lighter [47], FLASH [48], SAMtools [49], bwa mem [50] and Pilon [51]. A minimum contig length of 200 bp was used. Reference assemblies were produced for reference-based SNV calling. Briefly, the Oxford Nanopore Technology (ONT) MinION sequencer was used to generate long reads from 1D native barcoding libraries run on a R9.4 flow cell (ONT). Basecalling was carried out using Albacore v2.1.7 and v2.0.2 for the non-O157 and O157 datasets, respectively.

Combined with paired-end Illumina reads, hybrid assemblies were generated using Unicycler v0.4.6 [52] with default settings. The only exception was the O157:H7 reference genome (strain ECI-0907), which was constructed by assembling the Nanopore long reads using Canu [53], mapping paired-end Illumina reads back to the Canu assembly using Bowtie2 [54], removing low coverage contigs, polishing using Pilon [51], then circularizing using Circulator [55]. Assembly statistics are shown in Table 2.

In silico

Serotypes were confirmed using ECTyper v0.8.1 [56]. Cluster and phylogenetic analysis

Epidemiologically unrelated strains of the same serotype from public databases were included as outgroups, where available. No public O108:H8 data were available. Samples and related assemblies annotated as O108:H8 (i.e. SRX1983955, SRX1983925 and ERX406240) in the NCBI-SRA were typed by in silico serotyping as O108:H43, O108:H43 and O8:H8, respectively, and therefore excluded. The outgroup strain for O139:H19 (ERR439599) does not contain stx genes, but was used due to the absence of a suitable alternative. Tree visualizations were completed using FigTree v1.4.3 [57] or iTOL v3 [58].

Cluster and phylogenetic analysis

Epidemiologically unrelated strains of the same serotype from public databases were included as outgroups, where available. No public O108:H8 data were available. Samples and related assemblies annotated as O108:H8 (i.e. SRX1983955, SRX1983925 and ERX406240) in the NCBI-SRA were typed by in silico serotyping as O108:H43, O108:H43 and O8:H8, respectively, and therefore excluded. The outgroup strain for O139:H19 (ERR439599) does not contain stx genes, but was used due to the absence of a suitable alternative. Tree visualizations were completed using FigTree v1.4.3 [57] or iTOL v3 [58].

WgMLST

BioNumerics v7.6.2 E. coli wgMLST schema (17 380 loci) was applied for allele calls made from both assembly-free and assembly-based algorithms. Between 4000 and 6000 allele loci were used to reconstruct a dendrogram using the unweighted
pair group method with arithmetic mean (UPGMA) and the categorical (values) similarity coefficient. Minimum spanning trees (MSTs) were also reconstructed and partitions assigned based on a maximum of 10 allele differences between any two adjacent nodes. Branch lengths were scaled logarithmically to facilitate visualization. MST nodes were coloured by both cohort-year and individual cow IDs (for cows with ≥2 isolates).

**SNV analyses**

Panseq (commit ccb40f6 on Nov 2 2017) [59] was performed using the following parameters: ‘fragmentationSize’ = ‘500’, ‘percentIdentityCutoff’ = ‘99’, ‘coreGenomeThreshold’ = ‘[total no. of isolates in analysis]’, ‘cdhit’ = ‘1’. High sequence identity was used to minimize the incorporation of recombinant regions into the core genome [60]. Maximum-likelihood phylogeny was reconstructed using the concatenated SNP alignment using PhyML v3.0 [61], with default parameters: ‘substitution model selection criterion’ = ‘AIC’, ‘type of tree improvement’ = NNI, ‘branch support’ = ‘aLRT SH-like’.

SNVPhyl v1.0.1 [62] was performed using default parameters: minimum length for a repeat region = ‘150’ bp, minimum percent identity for a repeat region = ‘90’, minimum percent coverage of mapped reads = ‘80’, SNV abundance ratio = ‘0.75’, minimum mean mapping quality of reads for inclusion of a variant call = ‘30’, minimum read coverage for identification of variants = ‘15’, window size used for searching high-density SNV regions = ‘500’ bp, threshold of SNVs within the defined window size to flag high-density SNV regions = ‘2’. Hybrid assemblies from MinION long reads and Illumina short reads were used as references genomes for variant calling.

**Time-scaled phylogenies and evolutionary rates using Bayesian analysis**

TempEST v.1.5.1 [63] was used to assess whether there was sufficient temporal signal in each dataset using maximum-likelihood phylogenetic trees generated under the GTR (general time-reversible) + γ model in PhyML from the SNVPhyl SNP alignments. No reference nor outgroup strains were included in these phylogenies. A representative subset of the O157:H7 strains (n=36/106) was used to create a hetero-chronous dataset for analysis by TempEST, since the majority of strains were from 1996 and 1997. The ‘best root’ option was selected to generate the most ‘clock-like phylogeny’ based on maximizing the correlation of root-to-tip distance to sampling date. The root-to-tip genetic distances were plotted against the sampling dates and the resulting correlation coefficient and R² values were used to evaluate the strength of the relationship and the dispersion around the regression lines, respectively. Datasets were considered suitable for phylogenetic molecular clock analysis in beast v.1.8.4 [64] if the genetic divergence and sampling time exhibited a positive correlation.

Model construction for beast was completed in the program BEAUti v.1.8.4 [64] using the concatenated SNV from SNVPhyl alignments as input. Sampling dates were used for tip-dating calibration and the GTR + γ [4] nucleotide substitution model was used unless otherwise stated. All analyses were run for 10 million Markov chain Monte Carlo (MCMC) steps, sampling every 1000 steps. As per the work of Duchêne et al. [23], four model combinations were tested: strict molecular clock + constant size coalescent demographic model, strict molecular clock + Bayesian skyline demographic model, uncorrelated lognormal molecular clock + constant size coalescent demographic model, and uncorrelated lognormal molecular clock + Bayesian Skyline demographic model. For the Bayesian skyline demographic model, 10 groups and a piecewise-constant skyline model were employed. Marginal-likelihood estimation for each model was performed using path sampling (PS) and stepping-stone sampling (SS) implemented in BEAST. The model with the highest log marginal-likelihood was selected and used to generate two additional independent runs and combined in LogCombiner v1.8.2 [65], using a 10 % burn-in. Date randomization was performed in R (v3.6.2) using the RandomDates function in the ‘TipDatingBeast’ package and assessed as per the criteria described by Duchêne et al. [23]. Briefly, the clock.rate estimate from ten randomization replicates were compared with that from the run with accurate sampling dates. The strength of the temporal structure was classified based on the proportion of replicates for which the 95 % highest posterior density (HPD) intervals overlapped with that of the run with accurate sampling dates: 0 = ‘strong’, 0-0.5 = ‘moderate’, >0.5 = ‘low’ [23]. Tracer v1.7.0 [66] was used to assess the BEAST output (proper mixing/convergence, effective sample size ≥200) and a maximum clade credibility tree was generated using TreeAnnotator v1.8.4 [67] and visualized using FigTree v.1.4.3 [57].

Evolutionary rates were calculated as follows:

### Table 2. Quality statistics for hybrid assemblies of VTEC reference genomes

| Strain     | Serotype | No. of contigs | Largest contig (bp) | Total length (bp) | mol% G+C | N50, N75 |
|------------|----------|----------------|---------------------|-------------------|---------|---------|
| ECI-3359   | O6:H34   | 2              | 5 003 273           | 5 008 659         | 50.66   | 5 003 273 |
| ECI-2866   | O22:H8   | 8              | 5 026 381           | 5 180 090         | 50.79   | 5 026 381 |
| ECI-3462   | O108:H8  | 4              | 5 045 369           | 5 190 307         | 50.76   | 5 045 369 |
| ECI-3929   | O139:H19 | 3              | 4 964 669           | 5 147 626         | 50.79   | 4 964 669 |
| ECI-0907   | O157:H7  | 2              | 5 440 825           | 5 551 185         | 50.42   | 5 440 825 |
mean clock rate × (no. of SNV sites references genome core length) = substitutions (core site)−1 year−1

mean clock rate × no. of SNV sites = no. of SNVs per genome−1 per year−1

In silico detection of virulence factors, antibiotic-resistance genes and plasmid replicons

Using the ABRicate program v0.7 [68] and the following parameters, % coverage ≥60, % identity ≥85, all strains were interrogated for the presence of virulence factors (VFDB database 2597 sequences), antimicrobial-resistance (AMR) genes (ResFinder database 2228 sequences; CARD database 2153 sequences) and plasmid replicons (PlasmidFinder database 263 sequences) using databases last updated 27 August 2018. Additional virulence-associated genes were identified by VirulenceFinder v1.5 or v2.0 (% coverage ≥60, % identity ≥85) and genes not identified by ABRicate VFDB were reported [10].

Genome-wide association study (GWAS) of persistent versus sporadic serotypes

Persistent serotypes were defined as those isolated on a least ten different sampling dates overall and at least four sampling dates within each cohort-year. Sporadic serotypes were defined as those isolated on two sampling dates or fewer overall. The population structure of VTEC within this herd has been characterized previously [39]. This was used to assess for possible lineage effects that may confound the GWAS by ensuring that persistent and sporadic serotypes were not stratified by lineage; thus, leading to the identification of lineage-specific rather than phenotype-specific loci. Serotypes O22:H8 and O139:H19 were noted to form sub-clusters within a larger cluster [39].

A representative panel of persistent strains (n=13, four serotypes) and sporadic strains (n=11, eleven serotypes) (Table S3a) was annotated using Prokka v1.13 [69] and used to generate a pan-genome using Roary v3.13.0 [70] with the following parameters: -i [minimum percentage identity for BLASTP] = 95, -cd [percentage of isolates a gene must be in to be core] = 95, -s [don't split paralogs]. The presence/absence of 10 303 genes was used to assess the association with the ‘persistent’ trait using Scoary v1.6.16 [71], with the following parameters: -c [multiple comparison correction method] = BH (Benjamini–Hochberg), -p [P value cut-off] = 0.05. Significant genes were mapped to a reference E. coli genome (E. coli O91 strain RM7190, GenBank accession no. CP015244.1) using the BLAST Atlas function of GVView Server [72] with the following parameters: ‘minimum E value’ = 1e-10, ‘alignment length cut-off’ = 90, ‘per cent identity cut-off’ = 80. Putative gene functions were determined based on the reference genome protein annotations in addition to the Prokka annotations. Multiple core thresholds were tested (80, 85, 90, 95 and 99%) and while the number of core genes increased as the threshold was relaxed, this had no impact on the number or identity of the genes identified by the GWAS. Therefore, the default value of 99% was maintained. Using the output from Roary, IGRs were assessed for overrepresentation among persistent strains using Piggy with default parameters [73]. Lastly, ABRicate v0.7 [68] was used to screen the persistent and sporadic strains for virulence-associated genes, AMR genes and plasmid replicons, and Genome Fisher [74] was used to assess for statistically significant biases towards either group. The Benjamini–Hochberg correction for multiple comparisons was applied.

RESULTS
Genomic relatedness and population structure within persistent serotypes

wgMLST

Herd strains and epidemiologically unrelated strains were genetically distinct and temporal-cohort clustering was observed in all serotypes to varying degrees according to the MST method (Fig. 1). Tree topologies were supported by UPGMA (Figs S1–S3). A summary of the number of allele differences identified using both methods is reported in Table 3. The UPGMA allele differences describes the overall similarity of strains within a serotype and with epidemiologically unrelated strains. Given the cluster definition, the MST values describes whether serotype-specific strains can be grouped into one cluster or whether the relation is more diffuse.

O6:H34

Herd strains were genetically distinct from the human faecal outbreak strain, as well as a set of contemporaneous cattle isolates from Michigan, USA. Strains from cohort-year 3 clustered together and were within 0 to 3 alleles from a group of strains isolated from a single cow (ID 21101). Cohort-year 2 strains also clustered together, but were all isolated from the same individual. Isolates recovered from the same cow differed by 0 to 7 alleles using the MST method. High genetic similarity was also observed among the USA cattle strains, although metadata on herd characteristics were not available (Fig. 1a).

O22:H8

Herd strains were genetically distinct from all outgroup strains, except for a single stx2a-positive strain. The majority of strains from cohort-year 1 formed a distinct cluster, from which two clusters branched off: (i) cohort-year 3 stx2 (subtype c)-positive strains and (ii) cohort-year 2 and 3 stx2-negative strains (Fig. 1b). Isolates recovered from the same cow differed by 0 to 13 alleles using the MST method, except in two individuals where isolates with different stx toxin types were separated by >25 alleles. The same wgMLST profile was identified in two individuals.

O108:H8

Some temporal-cohort clustering was observed, with the majority of cohort-year 1 strains forming a distinct cluster. Sub-population within the largest cluster showed a trend towards delineation by cohort-year (Fig. 1c). Some clustering
Fig. 1. MSTs based on wgMLST profiles of persistent *E. coli* serotypes from cattle: (a) O6:H34, (b) O22:H8, (c) O108:H8, (d) O139:H19, (e) O157:H7. Main panels: distribution of genotypes among cohort-years relative to epidemiologically unrelated strains. Sub-panels: distribution of genotypes among individual cattle for which at least two isolates were obtained. Branch labels denote the number of allele differences. Isolates that share <10 allele differences with adjacent nodes are included in the partition (grey).
Table 3. Number of wgMLST allele differences and similarity values using UPGMA and MSTs. NA, Not applicable.

| Serotype | No. of allele differences | Similarity value* | No. of allele differences | Similarity value* | No. of clusters† | No. of allele differences |
|----------|---------------------------|-------------------|---------------------------|-------------------|------------------|---------------------------|
|          | Within herd               | Versus closest outgroup strain | Within herd               | Versus closest outgroup strain |                 |                           |
| O6:H34   | 10.6‡                     | 89.4‡              | 96.0                      | 4.0              | 1                | 75                        |
| O22:H8   | 28.7                      | 71.3               | 164.2                     | −64.2            | 2                | 87                        |
| O108:H8  | 42.8                      | 57.2               | NA                        | NA              | 6                | NA                        |
| O139:H19 | 106.0                     | −6.0               | >200                      | −100.0           | 1                | 400                       |
| O157:H7  | 11.7                      | 88.3               | 47.9§                     | 52.1§            | 1                | 371                       |

*Similarity value as calculated by BioNumerics v7.6.2 (−100 = 200 loci differences; 100 = identical wgMLST profiles).
†Cluster defined by 10 or fewer allele differences between any two adjacent nodes.
‡For comparison, within herd similarity of O6:H34 cattle isolates from the USA (n=23): 9.8 allele differences (90.2 similarity).
§Versus EDL933, 93.3 allele differences (6.7 similarity); versus Sakai, 99.1 allele differences (0.9 similarity).
||Versus EDL933, 72 allele differences; versus Sakai, 75 allele differences.

by individual cows was observed and the same wgMLST profile was identified in two individuals.

**O139:H19**
Herd strains were genetically distinct from the outgroup strain. Two sub-populations – one comprising primarily cohort-year 1 strains and the other primarily cohort-year 3 strains – were observed (Fig. 1d). Isolates from individual cows were generally more closely related and the same wgMLST profile was identified in two individuals.

**O157:H7**
Strains isolated between 1995 and 1998 were genetically distinct from the 2013 group of strains isolated within the same herd (≥59 alleles), and from the reference strains Sakai and EDL933 (>70 alleles). Within the 1990s group, strains were further stratified by year, with 0 to 10 alleles separating any two strains using the UPGMA method. The clustering of these strains appeared to be chronological, with genetic relatedness declining from 1995 strains to 1998 strains. Strains were generally not stratified by individual cows and multiple cows appear to have shed the same strain (Fig. 1e).

**SNV analyses**
The topologies of the maximum-likelihood phylogenies generated from pangenome-based (Panseq) and reference-based (SNVPhyl) SNV analyses were in general concordant with clustering by wgMLST, although some differences were observed among the three methods. Briefly, distinguishing features, such as temporal-cohort clustering and clustering by genetic feature such as the acquisition of stx2 in O22:H8, were maintained at these higher resolutions. Notable differences include: a break-up of the O6:H34 cohort-year 1 strains into separate clusters by SNVPhyl and less distinct clustering by year of O157:H7 strains by Panseq compared to wgMLST or SNVPhyl. SNV analysis did not generally allow for further resolution of clusters into epidemiologically relevant sub-clusters, e.g. cow-specific or season-specific, beyond those identified by wgMLST (Figs S4–S8).

**Estimated clock-like behaviour and core-genome evolutionary rates among persistent VTEC serotypes**
The most clock-like time-scaled phylogeny was generated using the 'best root' option in TempEST (Fig. 2). The estimated evolutionary rates (substitutions per site per year) and number of SNVs per genome per year are reported in Table 4. A negative correlation (coefﬁcient=−0.8477, R²=0.7186) was observed for a representative subset of the O157:H7 dataset (n=36) indicating no temporal signal, which was not resolved by the 'best-fitting root' option (Table 4). This precluded the dataset from further analysis using beast [63]. Temporal-cohort clustering was observed in all other serotypes, with genetic change progressing from the first to the last year of sampling (Table 4). The proportion of date randomization replicates for which the clock-rate HPDs overlapped with the clock-rate HPD estimated from using the accurate collection dates was 0 for all datasets, indicating strong temporal structure [23] (Fig. 3). Notable features from previous analyses, such as emergence of the stx2c-positive O22:H8 subgroup, were also observed (Fig. 2).

**Virulence factors, antibiotic resistance and plasmid gene content**
Several virulence-associated gene operons, including ent (enterobactin iron transport), fim (fimbriae) and yag (putative regulatory roles) were conserved across all serotypes, including in epidemiologically unrelated strains. Genes in the fep (ferric enterobactin transport) and gsp (cryptic general secretory pathway) operons were found in all four serotypes (O6:H34, O22:H8, O108:H8, O139:H19), but not in herd or reference O157:H7 strains. The AMR gene mdfA (E. coli multidrug transporter) was found in serotypes
Fig. 2. Bayesian inference of time-scaled phylogenies. (a) O6:H34. (b) O22:H8; pink highlighted strains, stx2c+ clade. (c) O108:H8. (d) O139:H19. Blue, cohort-year 1; green, cohort-year 2; red, cohort-year 3. Node and branch labels: posterior support. Node bars: height_95%_HPD.

Table 4. Summary of BEAST analysis of SNVPhyl SNV alignments. NA, Not applicable.

| Serotype | No. of strains* | Root-to-tip regression (correlation coefficient)† | Root-to-tip regression (R²)‡ | tMRCA‡ | Model§ | Mean clock rate | 95 % HPD interval | No. of SNV sites | No. of sites in reference core | Substitutions per site (core) per year¶ | SNVs per genome per year¶ |
|----------|----------------|-----------------------------------------------|-----------------------------|--------|--------|----------------|-----------------|----------------|-----------------------------|--------------------------|--------------------------|
| O6:H34   | 23             | 0.6934                                        | 0.4808                      | 2010.32| GTR + y, strict, skyline | 0.0756          | 0.0385, 0.1146  | 43              | 4588778                    | 7.08×10⁻⁷                | 3.25                     |
| O22:H8   | 30             | 0.9266                                        | 0.8586                      | 2009.87| GTR + y, strict, constant| 0.0712          | 0.0438, 0.1004  | 65              | 4671513                    | 9.91×10⁻⁷                | 4.63                     |
| O108:H8  | 40             | 0.8028                                        | 0.6445                      | 2011.32| GTR + y, strict, skyline | 0.0474          | 0.0312, 0.0633  | 140             | 4846133                    | 1.37×10⁻⁴                | 6.64                     |
| O139:H19 | 45             | 0.7530                                        | 0.5670                      | 2009.42| HKY + y, strict, constant| 0.0307          | 0.0244, 0.0392  | 196             | 4825351                    | 1.25×10⁻⁴                | 6.02                     |
| O157:H7  | 36#            | −0.8477                                       | 0.7186                      | NA     | Insufficient temporal signal for BEAST analysis |                |                 |                 |                            |                          |                          |

*Includes duplicate of reference strain; subtract one for the number of unique strains.†Using ‘best root’ option in TempEST for most clock-like phylogeny.‡tMRCA, estimated date of most recent common ancestor.§Nucleotide substitution model (GTR, HKY); clock type (strict, uncorrelated relaxed lognormal); tree prior (coalescent constant, coalescent Bayesian skyline).¶Mean clock rate × (no. of SNV sites/no. of sites in reference core).¶¶Mean clock rate × no. of SNV sites.

#Representative subset (n=36) of strains used to assess the temporal signal.
O6:H34, O108:H8 and O157:H7. No AMR genes were identified in O22:H8 or O139:H19 strains. With few exceptions, virulence gene profiles, antibiotic resistance and plasmid markers were conserved across all herd strains within each serotype (Table S2a–e).

**O6:H34**

Identical virulence gene profiles were observed for all herd and epidemiologically unrelated strains, with the exception of espX4, which was not detected in one herd strain. All herd and outgroup strains carried the antibiotic-resistance gene *mdf* and all strains from the USA bovine outgroup were also positive for other resistance-associated genes, including: *ant(3′)-Ia, aph(3′)-Ib, aph(6)-Id, dfrA1, floR, sul1, sul2 and tetA*. One plasmid replicon marker (ColRNAI) was detected in a single herd strain (Table S2a).

**O22:H8**

Identical virulence gene profiles were observed for most herd strains, except for a single strain in cohort-year 2 that carried *stx2c, cdtA, cdtB* and *cdtC*, and a subgroup of strains (n=10) in cohort-year 3 that carried *stx2a*. Some differences in gene content were observed when compared to epidemiologically unrelated strains. No antibiotic-resistance markers were identified, while one outgroup strain carried the following genes: *floR, strA, strB, sul2 and tetA*. Multiple plasmid replicons were detected in the majority of herd strains, including Col156, IncFIB(AP001918), IncFII(pHN7A8)_1_pHN7A8 and IncFII_1_pSFO; some of which were also identified in epidemiologically unrelated strains (Table S2b).

**O108:H8**

Identical virulence gene profiles were observed for all herd strains. All strains carried the antibiotic-resistance gene *mdf* and plasmid replicons ColRNAI and IncFIB(AP001918) (Table S2c).

**O139:H19**

Identical virulence gene profiles were observed for all herd strains. The following genes were only found in the herd strains and not in the outgroup strain, *cdtA, cdtB, cdtC, espP, stx1, stx2, iha, ehxA* and *subA*; whereas *astA* was only found in the outgroup strain. No antibiotic-resistance genes were identified. Multiple plasmid replicons were detected in all of the herd strains, including ColRNAI, IncB/O/K/Z, IncFIB(AP001918) and IncFII(pHN7A8) (Table S2d).

---

**Fig. 3.** Bayesian estimates of mean clock rates from date-randomization analysis. x-axes: 1–10, date randomization replicates; 11, accurate collection dates. y-axes: clock rate (substitutions per site). The strength of the temporal structure was rated based on the proportion of randomization replicates (n=10) for which the 95% HPDs overlapped with that estimated using accurate collection dates: 0, ‘strong’, 0–0.5, ‘moderate’, >0.5, ‘low’ [23].
O157:H7
Twenty virulence profiles were identified in herd strains of which one was dominant (n=64) and thirteen consisted of a single strain. Established enterohaemorrhagic E. coli virulence factors were present, including eae, toxB, hlyA, tir, stx1 and stx2. The antibiotic-resistance gene mdfA was present in all strains, including all outgroup and reference strains. Two plasmid replicons were detected among the herd strains, IncFIB(AP001918) and pEC4115_1, the former of which was also found in the outgroup and reference strains (Table S2e).

Potential acquisition of a phage-encoded stx2c in serotype O22:H8
Serotype-specific cluster analysis of O22:H8 showed that the stx2c-positive subgroup unique to cohort-year 3 formed a distinct clade within a larger clade comprising the majority of cohort-year 1 O22:H8 strains. A second clade was formed by stx2-negative strains from cohort-years 2 and 3 (Figs 1b, 2b, S1 and S5).

Accessory genetic features associated with VTEC persistence
Of the 176 virulence-associated, 2 AMR-associated and 12 plasmid replicon genes assessed by Genome Fisher, none were statistically overrepresented among the persistent group, nor among the sporadic group (Table S3c). Of the 10 303 genes assessed for association with the persistent phenotype, 14 were statistically overrepresented (P<0.005, Benjamini–Hochberg corrected for multiple comparisons) in persistent strains (n=13) relative to sporadic strains (n=11) (Fig. 4). The sensitivity and specificity of these genes were 100 % (13/13 persistent strains) and 90.9 to 100 % (1/11 to 0/11 sporadic strains), respectively. The six annotated genes encoded an outer-membrane porin protein OmpD (ompD_1), a double-stranded break reduction protein (rcbA), HTH-type transcriptional regulator YdeO (ydeO_2), a vitamin B12 transporter BtuB (btuB_1) and an arylsulfatase. The remaining nine were classified as hypothetical proteins (Table S3b). A high density of these genetic features mapped to a specific region of the genome of the reference strain E. coli O91 strain RM7190. Annotations within this region identified genes encoding an AraC family transcriptional regulator, cytochrome O ubiquinol oxidase, porin, sulfatase, vitamin B12/cobalamin outer-membrane transporter, transposase and several hypothetical proteins (Fig. 5). Of the 8659 IGRs assessed for association with the persistent phenotype, 11 were statistically overrepresented (P<0.005, Benjamini–Hochberg corrected for multiple comparisons) in persistent strains, with the sensitivity and specificity of these regions being 100 % (13/13 persistent strains) and 90.9 to 100 % (1/11 to 0/11 sporadic strains), respectively. These regions were flanked by sequences encoding hypothetical proteins, tRNA, and various structural and transport proteins (Table S3b).

DISCUSSION
Improved understanding of pathogen evolution and persistence in the cattle host continues to be important for farm-level mitigation of pathogenic E. coli [29, 75]. WGS analysis used in the surveillance of enteric pathogens enables unprecedented resolution of the genomic epidemiology of these organisms for foodborne illness investigations. WGS approaches applied to nine E. coli O157:H7 outbreaks in England from 2013 to 2017 afforded superior strain discrimination, case
ascertainment, and provided evidence for geographical origin and evolutionary context [9, 20]. In Listeria, it has improved cluster identification, facilitated retrospective trace-back of unsolved clusters and reduced faulty source attributions [76–79]. Within a distinct cattle population, we observed high genomic homogeneity within each persistently isolated serotype. Strains shared almost identical virulence gene, AMR gene and plasmid content, and were genetically distinct from epidemiologically unrelated strains. Given that at least three of these serotypes are known to cause human illness (O6:H34, O22:H8 and O157:H7), we evaluated the genomic relatedness of these serotypes in the context of a single farm source and determined the short-term evolutionary rate and genetic factors that may be indicative of persistence in cattle.

Within-serotype heterogeneity exceeded the provisional benchmark of ≤10 wgMLST alleles and varied between serotypes. Hence, even isolates of the same serotype from the same closed herd exhibit genomic differences over time. While these similarity thresholds are necessary for establishing inclusion/exclusion criteria for outbreak investigations, they may be too stringent for longer term surveillance. Isolates outside of proposed benchmarks of similarity but with sufficient phylogenetic or epidemiological evidence for relatedness should be considered for further investigation to avoid eliminating potentially credible sources of contamination [77]. Organism-, serotype- and even case-specific population genetics, and the spatial and temporal genomic variability inherent to different public-health contexts may be critical for resolving strain relatedness [19, 80] and identifying clades of clinical significance [81]. Serotypes that are genetically stable and those that are more variable may require different evaluation criteria [32]. Additionally, while the topology of the serotype-specific trees appears to be generally robust to SNP-based methods, the higher resolution does not appear to further resolve the clusters in terms of linking specific strains to outbreaks (e.g. relatedness within specific individuals or time periods) any further than by wgMLST. Therefore, higher resolution methods did not provide a better understanding of epidemiological data in this case. However, because the specific relationships between strains were not identical among the methods, there may be merit in employing different typing or phylogeny reconstruction methods to interpret the data.

Despite the relative persistence of clonal types, we observed cohort-temporal clustering and clonal turnover within serotypes. Particularly, serotype O157:H7 isolates clustered by year and displayed a logical progression in the number of wgMLST allele differences from 1995 to 1998. While similar trends were observed for non-O157 serotypes such as O6:H34 and O22:H8, in O139:H19 and O108:H8 multiple populations may have evolved, within which some cohort-temporal clustering was also observed. Persistence of predominant VTEC strains and clonal turnover on farms have been widely reported [32, 34, 82–86], although these studies have largely focused on O157:H7 and used lower resolution methods such as PFGE. Specifically, Cobbalt et al. [85] and Cobbold and Desmarchelier [83] identified predominant and farm-specific genotypes. Persistent VTEC strains may contribute to the diversification of the organism in cattle by providing a continual source of stx-phage and other transmissible virulence determinants, but also through continuous genetic change [82]. As such, altered genetic fingerprints may impact epidemiological investigations. Indeed, passage through cattle can result in spontaneous chromosomal deletions causing PFGE profile changes [87]. While the specific factors for persistence and subsequent genetic turnover are unknown for this particular herd, selective pressures including life events such as parturition, calving and weaning have been correlated with shedding of highly related E. coli O157:H7 at each stage [34]. Furthermore, we observed shared genotypes between individual cattle, suggesting that intra-herd dissemination is common. Horizontal transmission can contribute to the maintenance of E. coli O157:H7 and non-O157 in a herd [88–90], and may have contributed to the genetic heterogeneity we observed within these serotypes.

Continuous evolutionary changes within the genome were consistent with previously reported values for E. coli [15, 26–28]. Time-scale phylogenies were generally concordant with other methods and show a gradual accumulation of genetic change over time, sometimes in multiple subpopulations within a serotype. However, due to the low posterior supports (<0.5) for many of the nodes and branches, the clock rate estimates should be interpreted with caution. We observed the potential acquisition of the virulence-associated and phage-encoded stx2 gene in serotype O22:H8, which has
been associated previously with contamination of meat products [91], domestic animals [92], cattle [93] and haemolytic uremic syndrome patients [94]. The acquisition of stx-phages among a broad range of E. coli pathogroups may contribute to the genetic heterogeneity of VTEC in animals [32]. While the majority of these events appear to be transient in nature [95], in our study, carriage of stx2 in O22:H8 persisted throughout the year following acquisition. A 3 month sampling of the subsequent cohort (2015–2016) demonstrated that strains carrying stx2 had become the dominant O22:H8 genotype (data not shown). While it is possible a second O22:H8 genotype was introduced into the population, the otherwise identical repertoire of virulence gene, AMR gene and plasmid content, and the high genomic homogeneity shared with the stx2-negative subtype, suggest that this is the same clone that acquired a stx2-phage. Similarly, Geue et al. [89] found stx2-positive and -negative O26:H11 isolates within the same phylogenetic clusters.

Genes that were overrepresented in persistent serotypes included those that encoded OmpD/NmpC, RcbA/YdaC, YdeO and BtuB. The outer-membrane porin protein OmpD, most commonly found in Salmonella spp., and homologous with NmpC in E. coli K-12, is involved in ion transport. The presence of this gene has been correlated with Salmonella serovar host range [96]. Due to functional, regulatory and positional similarities between porin proteins in E. coli and Salmonella typhimurium [97], we can hypothesize that the overrepresentation of the gene in persistent serotypes may be associated with similar functions in E. coli. RcbA/YdaC maintains chromosome integrity by reducing the frequency of double-strand breaks [98]. Soo et al. [99] showed that over-expression of YdaC increased overall fitness and resistance to erythromycin in E. coli. YdeO is a transcriptional regulator involved in E. coli acid resistance [100], and BtuB is an outer-membrane transporter of vitamin B12, colicins and bacteriophages [101, 102]. The accumulation of persistence-associated determinants, which included a porin, sulfatase, vitamin B12/cobalamin outer-membrane transporter, transposase and hypothetical proteins, in a specific genomic region, suggests that these factors may be genetically linked. A large repertoire of colonization factors with affinity for the bovine intestinal epithelium and other biotic or abiotic surfaces, biofilm production and activation of stress fitness genes in bovine faeces may contribute to the persistence of certain E. coli [103].

The identification of genetic determinants for VTEC persistence may inform mitigation strategies at the farm level. For example, vaccination of cattle using colonization-associated proteins has reduced the level, prevalence and duration of E. coli O157:H7 shedding in experimental models [104].

In a similar investigation of long-term VTEC occurrence on cattle farms, four serotypes were persistent either among individual animals or herds (shedding for ≥4 months): O26:H11 (ST21/396/1705), O156:H25 (ST300/668), O165:H25 (ST119) and O182:H25 (ST300) [105]. We isolated O26:H11 (ST21) and O182:H25 (ST300) only sporadically [39], and the serotypes we identified as persistent were not isolated by Geue et al. [105]. Given that the latter study was conducted in Germany, it is possible that this reflects the endemcity of different VTEC serotypes in different geographical locations. In their case, subsequent PFGE cluster analysis of O165:H25, O26:H11 and O156:H25 demonstrated that population dynamics differed between serotypes. They observed differences in the tendency of certain serotypes to be farm- or animal-restricted, the potential loss of important colonization factors contributing to the loss of persistence, and the capacity of certain clones to persist within distinct temporal and spatial environments [89, 90, 106]. While the loss or low carriage of efa1 (E. coli colonization factor) was suggested to have resulted in the decline of a previously persistent serotype [106], efa1 was not present in any of the persistent serotypes we identified. Similar to our study, they found that isolates with identical genotypes were rarely isolated, demonstrating continuous genetic turnover within serotypes [90]. Barth et al. [36] identified certain virulence-associated genes that were more often found in persistent VTEC, including eae, efa-1-lipA, lpfA, tccP, espB, espI, nleA, nleB, nleC and stx1, while sporadic VTEC more often carried stx2 and toxB. In contrast, we identified lpfA in both persistent and sporadic serotypes [39], and all persistent serotypes carried stx2. Furthermore, the absence of significant biases of virulence factors, AMR or plasmid determinants toward the persistent or sporadic serotypes suggests that these elements may have limited impact on the persistence of VTEC in cattle. Conversely, the identification of IGRs and hypothetical proteins that are overrepresented in persistent strains may warrant further investigations into the roles of these elements. Therefore, the particular VTEC serotypes that persist, and the genetic determinants for this persistence, may reflect specific contamination sources, host dynamics and the relative fitness of serotypes within specific farm environments rather than be ascribed to a specific type of E. coli.

Several study limitations should be addressed. Although, the sampling of single age-cohorts within a closed herd was conducted out of convenience, it reduced the confounding effects of age, cattle rearing practices and the introduction of new cattle, effectively allowing us to observe the population and evolutionary dynamics of VTEC within a relatively controlled environment. Additionally, Rice et al. [107] found no effect of open versus closed herd management on the diversity of O157 subtypes on dairy farms, and non-bovine sources have been shown to also contribute to VTEC diversity within farms [84, 85]. Any discrepancies in topology observed between wgMLST, pan-genome and reference-derived SNP phylogenies may be the result of increased resolution afforded by SNP analyses, the additional assessment of IGRs and differences in the underlying algorithms used to estimate the true phylogeny. Ahrenfeldt et al. [108] estimated that only 50–73% of the true phylogeny could be recreated depending on which tools and settings were used. Therefore, in outbreak and surveillance investigations, it may be necessary to evaluate isolate relatedness using different tools to assess concordance with the underlying epidemiological evidence. The duration of sampling far exceeded that of typical outbreak investigations, while being comparatively short in the context...
of the global evolution of these pathogens. Duchène et al. [23] demonstrated that measurement of evolutionary rates on relatively short temporal scales captures deleterious mutants that will eventually be removed from the genome. Despite this, we were able to capture the potential clonal turnover of clinically relevant VTEC, including the acquisition of Shiga toxin 2, and demonstrate that for long-term surveillance, flexible interpretation of the criteria for isolate relatedness should be considered. The GWAS analysis was limited by the small sample size and because long-term persistence is a relatively difficult phenotype to observe unless populations are continuously monitored. Unfortunately, the sample size for the analysis was ultimately dictated by the limited number of ‘sporadic’ isolates available due to the overwhelming dominance of the four persistent serotypes among the isolates. Future studies may be facilitated by harnessing larger datasets collected from routine surveillance to identify ‘persistent’ serotypes using a population of interest e.g. serotype persistence based on clinical outcomes, or regional, temporal, commodity- or source-based associations. Additionally, more sophisticated methods for GWAS may address the issue of sample size imbalance [109]. Furthermore, we focused on factors that were biased towards persistence, whereas factors that are overrepresented in sporadic strains may also be of interest. Lastly, despite being from the same herd, the O157 and non-O157 datasets should be compared with caution due to differences in study design, including isolation method, cohort selection and collection years.

In summary, we observed high genomic homogeneity among bovine isolates from a single farm source within clinically relevant VTEC serotypes. While the number of allele differences are similar to current provisional thresholds used for clustering human strains associated with single source outbreaks, these thresholds may need to be re-evaluated for surveillance applications and may differ between serotypes. High genomic homogeneity, temporal-cohort clustering and observed changes to the genome at rates consistent with previous estimates for E. coli provide evidence of clonal turnover of dominant serotypes within cattle. Differences in evolutionary rates between serotypes should be considered when evaluating strain similarity for long-term surveillance. These changes included the acquisition of a phage-encoded virulence factor, demonstrating the capacity of the bovine environment to facilitate the transfer of virulence-associated elements into clinically relevant serotypes. Lastly, herd persistence may be characterized by genetic features involved in small molecule transport, chromosome repair, AMR, acid resistance and general fitness. These results may have implications for cluster identification during outbreak investigation, pathogen surveillance and informing farm-level mitigation efforts for VTEC.

Acknowledgements
The authors thank the Core Genomics Facility at the National Microbiology Laboratory (Winnipeg, Manitoba, Canada) for WGS of some of the study strains. The authors also thank James Robertson (National Microbiology Laboratory, Guelph, Ontario, Canada) for his expertise on Nanopore sequencing and analysis.

Author contributions
Conceptualization: L.Y.R.W., C.R.L., V.P.J.G. Data curation: L.Y.R.W., C.C.J., C.R.L. Formal analysis: L.Y.R.W., C.R.L. Funding acquisition: C.C.J., C.R.L., R.P.J., K.Z., V.P.J.G. Investigation: L.Y.R.W., C.C.J., C.R.L. Methodology: L.Y.R.W., C.C.J., R.P.J., K.Z., V.P.J.G. Project administration: V.P.J.G. Resources: C.R.L., R.P.J., K.Z., V.P.J.G. Software: C.R.L. Supervision: V.P.J.G. Validation: L.Y.R.W. Visualization: L.Y.R.W. Writing – original draft: L.Y.R.W. Writing – review and editing: L.Y.R.W., C.C.J., C.R.L., R.P.J., K.Z., V.P.J.G.

Conflicts of interest
The authors declare that there are no conflicts of interest.

Ethical statement
This study was carried out in accordance with the principles contained in the Guide to the Care and Use of Experimental Animals, vols I and II, by the Canadian Council on Animal Care (CCAC) and followed CCAC guidelines. The protocol was approved by the CFIA-NCAD Animal Care Committee (protocol no. 1201).

Data Bibliography
1. Wang LYR, Jokinen CC, Laing CR, Johnson RP, Ziebell K, Gannon VPJ. NCBI Sequence Read Archive (SRA) PRJNA565946 (2019).
2. Centers for Disease Control and Prevention (CDC). NCBI Sequence Read Archive (SRA) SRR2038679 (2015).
3. US Food and Drug Administration (FDA). NCBI Sequence Read Archive (SRA) SRR3133016 (2015).
4. US Food and Drug Administration (FDA). NCBI Sequence Read Archive (SRA) SRR3929485, SRR4263660 (2016).
5. Public Health England (PHE). NCBI Sequence Read Archive (SRA) SRR4184713 (2016).
6. Canadian Food Inspection Agency (CFIA). NCBI Sequence Read Archive (SRA) SRR6061349, SRR6061353 (2017).
7. National Health Service (NHS) Lothian. NCBI Sequence Read Archive (SRA) SRR6321291 (2017).
8. Wellcome Trust Sanger Institute. NCBI Sequence Read Archive (SRA) ERR439599 (2014).
9. US Food and Drug Administration (FDA). GenBank GCA_002515685.1, GCA_002514685.1, GCA_002514585.1, GCA_002514685.1, GCA_002515445.1, GCA_002515435.1, GCA_002515435.1, GCA_002515435.1, GCA_002514957.1, GCA_002514957.1, GCA_002514957.1, GCA_002514957.1, GCA_002514957.1, GCA_002514957.1 (2018).
10. Latif H. University of California San Diego. GenBank NZ_CP008957.1 (strain EDL933) (2014).
11. Juenemann S. ATCC. GenBank NC_002695.2 (strain Sakai) (2013).
12. Centers for Disease Control and Prevention (CDC). NCBI Sequence Read Archive (SRA) PRJNA565946 (2019).

References
1. Bettelheim KA. The non-O157 Shiga-toxigenic (verocytotoxigenic) Escherichia coli; under-rated pathogens. Crit Rev Microbiol 2007;33:67–87.
2. Dallman TJ, Byrne L, Ashton PM, Cowley LA, Perry NT et al. Whole-genome sequencing for national surveillance of Shiga toxin-producing Escherichia coli O157. Clin Infect Dis 2015;61:305–312.
3. Chattaway MA, Dallman TJ, Gentile A, Wright MJ, Long SE et al. Whole genome sequencing for public health surveillance of Shiga toxin-producing Escherichia coli other than serogroup O157. Front Microbiol 2016;7:258.
4. Lindsey RL, Pouseele H, Chen JC, Stockbine NA, Carleton HA. Implementation of whole genome sequencing (WGS) for identification and characterization of Shiga toxin-producing Escherichia coli (STEC) in the United States. Front Microbiol 2016;7:766.

5. Holmes A, Dallman TJ, Shabaan S, Hanson M, Allison L. Validation of whole-genome sequencing for identification and characterization of Shiga toxin-producing Escherichia coli to produce standardized data to enable data sharing. J Clin Microbiol 2018;56:e01388-17.

6. Lüth S, Kleta S, Al Dahouk S. Whole genome sequencing as a typing tool for foodborne pathogens like Listeria monocytogenes – the way towards global harmonisation and data exchange. Trends Food Sci Technol 2018;73:67–75.

7. Grad YH, Lipsitch M, Feldgarden M, Arachchi HM, Cerqueira GC et al. Genomic epidemiology of the Escherichia coli O104:H4 outbreaks in Europe, 2011. Proc Natl Acad Sci USA 2012;109:3065–3070.

8. Rumore J, Tscherter L, Kearney A, Kandar R, McCormick R et al. Evaluation of whole-genome sequencing for outbreak detection of verotoxigenic Escherichia coli O157:H7 from the Canadian perspective. BMC Genomics 2018;19:870.

9. Jenkins C, Dallman TJ, Grant KA. Impact of whole genome sequencing on the investigation of food-borne outbreaks of Shiga toxin-producing Escherichia coli serogroup O157:H7, England, 2013 to 2017. Euro Surveill 2019;24:1800346.

10. Joensen KG, Scheutz F, Lund O, Hasman H, Kaas RS et al. Real-time whole-genome sequencing for routine typing, surveillance, and outbreak detection of verotoxigenic Escherichia coli. J Clin Microbiol 2014;52:1501–1510.

11. Holmes A, Allison L, Ward M, Dallman TJ, Clark R et al. Utility of whole-genome sequencing of Escherichia coli O157 for outbreak detection and epidemiological surveillance. J Clin Microbiol 2015;53:3565–3573.

12. Mikhail AFW, Jenkins C, Dallman TJ, Inns T, Douglas A et al. An outbreak of Shiga toxin-producing Escherichia coli O157:H7 associated with contaminated salad leaves: epidemiological, genomic and food trace back investigations. Epidemiol Infect 2018;146:187–196.

13. Orsi RH, Borowsky ML, Lauer P, Young SK, Nusbaum C et al. Short-term genome evolution of Listeria monocytogenes in a non-controlled environment. BMC Genomics 2008;9:539.

14. Haugum K, Johansen J, Gabrielsen C, Brandal LT, Bergh K et al. Comparative genomics to delineate pathogenic potential in non-O157 Shiga toxin-producing Escherichia coli (STEC) from patients with and without haemolytic uraemic syndrome (HUS) in Norway. PLoS One 2014;9:e117788.

15. Dallman TJ, Ashton PM, Byrne L, Perry NT, Petrovska L et al. Applying phylogenomics to understand the emergence of Shiga-toxin-producing Escherichia coli O157:H7 strains causing severe human disease in the UK. Microb Genom 2015;1:e000029.

16. Dekker JP, Frank KM. Next-generation epidemiology: using real-time core genome multilocus sequence typing to support infection control policy. J Clin Microbiol 2016;54:2850–2853.

17. Roer L, Hansen F, Thomsen MCF, Knudsen JD, Hansen DS et al. WGS-based surveillance of third-generation cephalosporin-resistant Escherichia coli from bloodstream infections in Denmark. J Antimicrob Chemother 2017;72:1922–1929.

18. Underwood AP, Dallman T, Thomson NR, Williams M, Harker K et al. Public health value of next-generation DNA sequencing of enterohemorrhagic Escherichia coli isolates from an outbreak. J Clin Microbiol 2013;51:232–237.

19. Schürch AC, Arredondo-Alonso S, Willems RJJ, Goering RV. Whole genome sequencing options for bacterial strain typing and epidemiologic analysis based on single nucleotide polymorphism versus gene-by-gene–based approaches. Clin Microbiol Infect 2018;24:350–356.

20. Butcher H, Elson R, Chattaway MA, Featherstone CA, Willis C et al. Whole genome sequencing improved case ascertainment in an outbreak of Shiga toxin-producing Escherichia coli O157 associated with raw drinking milk. Epidemiol Infect 2016;144:2812–2823.

21. Caprioli A, Morabito S, Brugère H, Oswald E. Enterohaemorrhagic Escherichia coli: emerging issues on virulence and modes of transmission. Vet Res 2005;36:289–311.

22. Bettelheim KA, Goldwater PN. Serotypes of non-O157 shigatoxigenic Escherichia coli (STEC). Adv Microbiol 2014;4:377–389.

23. Duchène S, Holt KE, Weill F, Le Hello S, Hawkey J et al. Genomic-scale rates of evolutionary change in bacteria. Microb Genom 2016;2:e000094.

24. Casali N, Broda A, Harris SR, Parkhill J et al. Whole genome sequence analysis of a large isoniazid-resistant tuberculosis outbreak in London: a retrospective observational study. PLoS Med 2016;13:e1002137.

25. Kennemann L, Didelot X, Aebischer T, Kuhn S, Drescher B et al. Helicobacter pylori genome evolution during human infection. Proc Natl Acad Sci USA 2011;108:5033–5038.

26. von Mentzer A, Conroy TR, Wieler LH, Semmler T, Iguchi A et al. Identification of entero-oxicogenic Escherichia coli (ETEC) clades with long-term global distribution. Nat Genet 2014;46:1321–1326.

27. Holt KE, Thieu Ngia TV, Pham Thanh D, Vinh H, Kim DW et al. Tracking the establishment of local endemic populations of an emergent enteric pathogen. Proc Natl Acad Sci USA 2013;110:17522–17527.

28. Stoesser N, Sheppard AE, Pankhurst L, De Maio N, Moore CE et al. Evolutionary history of the global emergence of the Escherichia coli epidemic clone ST131. mBio 2016;7:e02162.

29. Larsen MH, Dalmaso M, Ingmer H, Langsrud S, Malakauskas M et al. Persistence of foodborne pathogens and their control in primary and secondary food production chains. Food Control 2014;44:92–109.

30. Stasiewicz MJ, Oliver HF, Wiedmann M, den Bakker HC. Whole-genome sequencing allows for improved identification of persistent Listeria monocytogenes in food-associated environments. Appl Environ Microbiol 2015;81:6024–6037.

31. Malley TV, Butts J, Wiedmann M. Seek and destroy process: Listeria monocytogenes process controls in the ready-to-eat meat and poultry industry. J Food Prot 2015;78:436–445.

32. Beutin L, Geier D, Zimmermann S, Aleksic S, Gillespie HA et al. Epidemiological relatedness and clonal types of natural populations of Escherichia coli strains producing Shiga toxins in separate populations of cattle and sheep. Appl Environ Microbiol 1999;63:2175–2180.

33. Shere JA, Bartlett KJ, Kaspar CW. Longitudinal study of Escherichia coli O157:H7 dissemination on four dairy farms in Wisconsin. Appl Environ Microbiol 1998;64:1390–1399.

34. Gannon VPJ, Graham TA, King R, Michel P, Read S et al. Escherichia coli O157:H7 infection in cows and calves in a beef cattle herd in Alberta, Canada. Epidemiol Infect 2002;129:163–172.

35. Carlson BA, Nightingale KK, Mason OL, Ruby JR, Chat W et al. Escherichia coli O157:H7 strains that persist in feedlot cattle are genetically related and demonstrate an enhanced ability to adhere to intestinal epithelial cells. Appl Environ Microbiol 2009;75:5927–5937.

36. Barth SA, Menge C, Eichhorn I, Semmler T, Wieler LH et al. The accessory genome of Shiga toxin-producing Escherichia coli defines a persistent colonization type in cattle. Appl Environ Microbiol 2016;82:5455–5464.

37. Buchanan CJ, Webb AL, Mutschall SK, Kuczkwiewicz P, Barker DOR et al. A genome-wide association study to identify diagnostic markers for human pathogenic Campylobacter jejuni strains. Front Microbiol 2017;8:1224.

38. Lewis BB, Carter RA, Ling L, Leiner I, Taur Y et al. Pathogenicity locus, core genome, and accessory gene contributions to Clostridium difficile sequience. mBio 2017;8:e00885-17.

39. Wang LYR, Jokinen CC, Laing CR, Johnson RP, Ziebell K et al. Multi-year persistence of verotoxigenic Escherichia coli (VTEC)
in a closed Canadian beef herd: a cohort study. Front Microbiol 2018;9:2040.

40. Matthews TC, Bristow FR, Griffiths EJ, Petkau A, Adam J et al. The integrated rapid infectious disease analysis (IRIDA) platform. bioRxiv 2018:381830.

41. Seemann T, Kwong J, Gladman S, da Silva AG. Shovill, version 1.1.0, 2018. https://github.com/tseemann/shovill

42. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol 2012;19:455–477.

43. Ondov BD, Treangen TJ, Mertens P, Mallonee AB, Bergman NH et al. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol 2016;17:132.

44. Li H. Seqtk: toolkit for processing sequences in FASTA/Q formats; 2012.

45. Deorowicz S, Kokot M, Grabowski S, Debudaj-Grabsz A, KMC 2: fast and resource-frugal k-mer counting. Bioinformatics 2015;31:1569–1576.

46. Bolger AM, Lohse M, Usadel B. Trimmmomatic: a flexible trimmer for illumina sequence data. Bioinformatics 2014;30:2114–2120.

47. Song L, Florea L, Langmead B. Lighter: fast and memory-efficient sequence correction error correction without counting. Genome Biol 2014;15:509.

48. Magoč T, Salzberg SL. FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics 2011;27:2957–2963.

49. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J et al. The sequence alignment MAP format and SAMtools. Bioinformatics 2009;25:2078–2079.

50. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM; 2013. https://arxiv.org/abs/1303.3997

51. Walker BJ, Abeel T, Shea T, Priest M, Abouelbel A et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One 2014;9:e112963.

52. Wick RR, Judd LM, Gorrie CL, Holt KE. Unicyclic: resolving bacterial genome assemblies from short and long sequencing reads. PLoS Comput Biol 2017;13:e1005595.

53. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH et al. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res 2017;27:722–736.

54. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods 2012;9:357–359.

55. Hunt M, Silva ND, Otto TD, Parkhill J, Keane JA et al. Circlator: automated circularization of genome assemblies using long sequencing reads. Genome Biol 2015;16:294.

56. Le KK, Whiteside MD, Hopkins JE, Gannon VPJ, Laing CR. Sply: an integrated graph database for real-time prediction of bacterial phenotypes and downstream comparative analyses. Database 2018;2018:baaye08.

57. Rambaut A. FigTree, version 1.4.3; 2016.

58. Letunic I, Bork P. Interactive tree of life (iTOL): an online tool for phylogenetic tree display and annotation. Bioinformatics 2007;23:127–128.

59. Laing C, Buchanan C, Taboada EN, Zhang Y, Kropinski A et al. Pan-genome sequence analysis using Panaqc: an online tool for the rapid analysis of core and accessory genomic regions. BMC Bioinformatics 2010;11:461.

60. Franz E, Rotariu O, Lopes BS, MacRae M, Bono JL et al. Phylogeographic analysis reveals multiple international transmission events have driven the global emergence of Escherichia coli O157:H7. Clin Infect Dis 2018;69:428–437.

61. Guindon S, Lethiec F, Duroux P, Gascuel O. PHYML online – a web server for fast maximum likelihood phylogenetic inference. Nucleic Acids Res 2005;33:W557–559.

62. Petkau A, Mabon P, Sieffert C, Knox NC, Cabral J et al. SNVPhyt: a single nucleotide variant phylogenomics pipeline for microbial genomic epidemiology. Microb Genom 2017;3:e000116.

63. Rambaut A, Lam TT, Max Carvalho L, Pybus OG. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). Virus Evol 2016;2:vew007.

64. Drummond AJ, Suchard MA, Xie D, Rambaut A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol Biol Evol 2012;29:1969–1973.

65. Rambaut A, Drummond AJ. LogCombiner, version 1.8.2; 2015. http://beast.bio.ed.ac.uk

66. Rambaut A, Suchard MA, Xie W, Drummond AJ. Tracer: MCMC trace analysis tool, version 1.7.0; 2013. http://beast.bio.ed.ac.uk

67. Rambaut A, Drummond AJ. TreeAnnotator: MCMC output analysis, version 1.8.4; 2015. http://beast.bio.ed.ac.uk

68. Seemann T. ABriCate: mass screening of contigs for antimicrobial resistance or virulence genes, version 0.7; 2017. https://github.com/tseemann/abriicate

69. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics 2014;30:2068–2069.

70. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S et al. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics 2015;31:3691–3693.

71. Bryniøsrud O, Bohlin J, Scheffer L, Eldholm V. Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. Genome Biol 2016;17:238.

72. Stothard P, Grant JR, Van Domselaar G. Visualizing and comparing circular genomes using the CGView family of tools. Brief Bioinform 2019;20:1576–1582.

73. Thorpe HA, Bayliss SC, Sheppard SK, Fell EJ. Piggy: a rapid, large-scale pan-genome analysis tool for intergenic regions in bacteria. Gigascience 2017;8:gty015.

74. Kuczkieiwicz P. Genome Fisher; 2013. https://bitbucket.org/peterk87/genomefisher

75. Soon JM, Chadd SA, Baines RN. Escherichia coli 0157:H7 in beef cattle: on farm contamination and pre-slaughter control methods. Anim Health Res Rev 2011;12:197–211.

76. CDC. Multistate outbreak of listeriosis linked to soft cheeses distributed by Karoun Dairies, Inc. (final update). Atlanta, GA: CDC; 2015. https://www.cdc.gov/listeria/outbreaks/soft-cheeses-09-15/

77. Schmid D, Allerberger F, Huhulescu S, Pietzka A, Mac C et al. Whole genome sequencing as a tool to investigate a cluster of seven cases of listeriosis in Austria and Germany, 2011–2013. Clin Microbiol Infect 2014;20:431–436.

78. Jackson BR, Tarr C, Strain E, Jackson KA, Conrad A et al. Implementation of nationwide real-time whole-genome sequencing to enhance listeriosis outbreak detection and investigation. Clin Infect Dis 2016;63:380–386.

79. Kleta S, Hammerl JA, Dieckmann R, Malorny B, Borowiak M et al. Molecular tracing to find source of protracted invasive listeriosis outbreak, southern Germany, 2012–2016. Emerg Infect Dis 2017;23:1680–1683.

80. Goering RV, Köck R, Grundmann H, Werner G, Friedrich AW et al. From theory to practice: molecular strain typing for the clinical and public health setting. Euro Surveill 2013;18:20383.

81. Wirth T, Falush D, Lan R, Colles F, Mensa P et al. Sex and virulence in Escherichia coli: an evolutionary perspective. Mol Microbiol 2006;60:1136–1151.

82. Akiba M, Sameshima T, Nakazawa M. Clonal turnover of enterohemorrhagic Escherichia coli O157:H7 in experimentally infected cattle. FEMS Microbiol Lett 2000;184:79–83.

83. Cobbold R, Desmarchelier P. Characterisation and clonal relationships of Shiga-toxigenic Escherichia coli (STEC) isolated from Australian dairy cattle. Vet Microbiol 2001;79:323–335.

84. Schouten JM, Graat EAM, Franken K, van de Giessen AW, van der Zwaluw WK et al. A longitudinal study of Escherichia coli O157
in cattle of a Dutch dairy farm and in the farm environment. Vet Microbiol 2005;107:193–204.

85. Cobbaut K, Houf K, Boyen F, Haesebrock F, De Zutter L. Genotyping and antimicrobial resistance patterns of Escherichia coli O157 originating from cattle farms. Foodborne Pathog Dis 2011;8:719–724.

86. Joris M-A, Verstraete K, De Reu K, De Zutter L. Longitudinal follow-up of the persistence and dissemination of EHEC on cattle farms in Belgium. Foodborne Pathog Dis 2013;10:295–301.

87. Yoshii N, Ogura Y, Hayashi T, Ajiro T, Sameshima T et al. Pulsed-field gel electrophoresis profile changes resulting from spontaneous chromosomal deletions in enterohemorrhagic Escherichia coli O157:H7 during passage in cattle. Appl Environ Microbiol 2009;75:5722–5726.

88. Faith NG, Shere JA, Brosch R, Arnold KW, Ansay SE et al. Prevalence and clonal nature of Escherichia coli O157:H7 on dairy farms in Wisconsin. Appl Environ Microbiol 1996;62:1519–1525.

89. Geue L, Klare S, Schnick C, Mintel B, Meyer K et al. Analysis of the clonal relationship of serotype O26:H11 enterohemorrhagic Escherichia coli isolates from cattle. Appl Environ Microbiol 2009;75:6947–6953.

90. Geue L, Schaars S, Mintel B, Conraths FJ, Müller E et al. Rapid microarray-based genotyping of enterohemorrhagic Escherichia coli serotype O156:H25/F-11/minis from cattle and clonal analysis. J Appl Microbiol 2010;76:5510–5523.

91. Cadon J, Bustamante AV, González J, Sanso AM. Genetic relatedness and novel sequence types of non-O157 Shiga toxin-producing Escherichia coli isolates from Argentina. Front Cell Infect Microbiol 2016;6:93.

92. Bentancor A, Rumi MV, Gentilini MV, Sardoy C, Grisio K et al. Shiga toxin-producing and attaching and effacing Escherichia coli in cats and dogs in a high hemolytic uremic syndrome incidence region in Argentina. FEMS Microbiol Lett 2007;267:251–256.

93. Menrath A, Wieler LH, Heidemann K, Semmler T, Fruth A et al. Shiga toxin producing Escherichia coli; identification of non-O157:H7 super-sheathing cows and related risk factors. Gut Pathog 2010;2:7.

94. Blanco J, Blanco M, Blanco JE, Mora A, González EA et al. Verotoxin-producing Escherichia coli in Spain: prevalence, serotypes, and virulence genes of O157:H7 and non-O157 VTEC in ruminants, raw beef products, and humans. Exp Biol Med 2003;228:345–351.

95. Tozzi R, Grande L, Michelacci V, Ranieri P, Maugliani A et al. Shiga toxin-converting phages and the emergence of new pathogenic Escherichia coli; a world in motion. Front Cell Infect Microbiol 2014;4:80.

96. Santiago CA, Toro CS, Bucarey SA, Mora GC. A chromosomal region surrounding the ompD porin gene marks a genetic difference between Salmonella typhi and the majority of Salmonella serovars. Microbiology 2001;147:1897–1907.

97. Lee DR, Schnaitman CA. Comparison of outer membrane porin proteins produced by Escherichia coli and Salmonella typhimurium. J Bacteriol 1980;142:1019–1022.

98. Felczak MM, Kaguni JM. The rcbA gene product reduces spontaneous and induced chromosome breaks in Escherichia coli. J Bacteriol 2012;194:2152–2164.

99. Soo VWC, Hanson-Manful P, Patrick WM. Artificial gene amplification reveals an abundance of promiscuous resistance determinants in Escherichia coli. Proc Natl Acad Sci USA 2011;108:1484–1489.

100. Masuda N, Church GM. Regulatory network of acid resistance genes in Escherichia coli. Mol Microbiol 2003;48:697–712.

101. Di Masì DR, White J, Schnaitman CA, Bradbeer C. Transport of vitamin B12 in Escherichia coli: common receptor sites for vitamin B12 and the E coli colicins on the outer membrane of the cell envelope. J Bacteriol 1973;115:504–513.

102. Bradbeer C, Woodrow ML. Transport of vitamin B12 in Escherichia coli: energy dependence. J Bacteriol 1976;128:99–104.

103. Segura A, Auffret P, Bibb D, Bertoni M, Durand A et al. Factors involved in the persistence of a Shiga toxin-producing Escherichia coli O157:H7 strain in bovine feces and gastro-intestinal content. Front Microbiol 2018;9:375.

104. Potter AA, Klashinsky SI, Li Y, Frew E, Townsend H et al. Decreased shedding of Escherichia coli O157:H7 by cattle following vaccination with type III secreted proteins. Vaccine 2004;22:362–369.

105. Geue L, Segura-Alvarez M, Conraths FJ, Kuczious T, Bockemühl J et al. A long-term study on the prevalence of Shiga toxin-producing Escherichia coli STEC on four German cattle farms. Epidemiol Infect 2002;129:173–185.

106. Geue L, Selhorst T, Schnick C, Mintel B, Conraths FJ. Analysis of the clonal relationship of Shiga toxin-producing Escherichia coli serogroup O165:H25 isolated from cattle. Appl Environ Microbiol 2006;72:2254–2259.

107. Rice DH, McMenamin KM, Pritchett LC, Hancock DD, Besser TE. Genetic subtyping of Escherichia coli O157 isolates from Pacific Northwest USA cattle farms. Epidemiol Infect 1999;122:479–484.

108. Ahrenfeldt J, Skaarup C, Hasman H, Pedersen AG, Aarestrup FM et al. Bacterial whole genome-based phylogeny: construction of a new benchmarking dataset and assessment of some existing methods. BMC Genomics 2017;18:19.

109. Zhou W, Nielsen JB, Fritsche LG, Dey R, Gabrielsen ME et al. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. Nat Genet 2018;50:1335–1341.