A standardised method for determining *Escherichia coli* O157:H7 strain relatedness using whole genome sequencing or virulence gene profiling is not yet established. We sought to assess the capacity of either high-throughput polymerase chain reaction (PCR) of 49 virulence genes, core-genome single nt variants (SNVs) or *k*-mer clustering to discriminate between outbreak-associated and sporadic *E. coli* O157:H7 isolates. Three outbreaks and multiple sporadic isolates from the province of Alberta, Canada were included in the study. Two of the outbreaks occurred concurrently in 2014 and one occurred in 2012. Pulsed-field gel electrophoresis (PFGE) and multilocus variable-number tandem repeat analysis (MLVA) were employed as comparator typing methods. The virulence gene profiles of isolates from the 2012 and 2014 Alberta outbreak events and contemporary sporadic isolates were mostly identical; therefore the set of virulence genes chosen in this study were not discriminatory enough to distinguish between outbreak clusters. Concordant with PFGE and MLVA results, core genome SNV and *k*-mer phylogenies clustered isolates from the 2012 and 2014 outbreaks as distinct events. *k*-mer phylogenies demonstrated increased discriminatory power compared with core SNV phylogenies. Prior to the widespread implementation of whole genome sequencing for routine public health use, issues surrounding cost, technical expertise, software standardisation, and data sharing/comparisons must be addressed.

**Introduction**

Shiga toxin-producing *Escherichia coli* (STEC) consisting of O157 and non-O157 serogroups are a major public health concern. Cattle and other ruminants are natural reservoirs for STEC organisms, shedding the organisms in their faeces, which can cause food and/or water contamination [1]. Consumption of contaminated meat, dairy products, vegetables/fruit, water, contact with animals [1] and person-to-person transmission [2] have all been associated with STEC infections. Infection may be asymptomatic or can cause gastrointestinal symptoms, including mild diarrhoea to haemorrhagic colitis [3]. In five to 20% of the infected patients post-diarrhoea haemolytic uraemic syndrome (HUS) occurs, which is characterised by haemolytic anaemia, thrombocytopenia and kidney injury or failure [2,3]. Paediatric and elderly patients are at greatest risk for developing systemic STEC complications, which are not limited to HUS and can include cardiac, central nervous system, pancreatic, and pulmonary complications [3-5]. Shiga toxins (Stx1 and Stx2) are the major virulence determinants responsible for symptoms associated with both haemorrhagic colitis and systemic infections [5].

Due to the public health importance of STEC infections, epidemiological and molecular surveillance systems are essential for early outbreak detection. In recent years, rapid advancements in the use of molecular typing methods have improved STEC surveillance and outbreak detection. The application of these tools helps to identify disease clusters, refine outbreak case definitions, facilitate case finding, and link human cases to environmental sources. In order to achieve these outcomes, molecular typing assays must possess the discriminatory power required to distinguish between related and nonrelated bacterial isolates, have high reproducibility, and be easy to perform. Furthermore,
In Alberta, Canada, all *E. coli* O157:H7 are routinely typed by pulsed-field gel electrophoresis (PFGE) and multilocus variable-number tandem repeat analysis (MLVA) under surveillance practices. In July 2014, a unique cluster of *E. coli* O157:H7 was identified by PFGE with concordant MLVA analysis. The subsequent molecular and epidemiological investigation revealed that the cluster was associated with one of the largest human *E. coli* O157:H7 outbreaks in Canada since the implementation of PulseNet Canada (PNC; a national molecular subtyping network for food-borne disease surveillance) in 2000 (Linda Chui, PNC internal communications). To investigate the relatedness of isolates, whole genome sequencing (WGS) and virulence gene profiling were performed separately in real-time with concomitant analysis. The objectives of this study were twofold: (i) We sought to determine the relatedness of this large outbreak event to a concurrent, albeit smaller outbreak as well as to all sporadic cases occurring in the summer of 2014 in Alberta. In addition, a representative panel of isolates from a socially and economically significant, 2012 beef-associated outbreak was included for comparison. (ii) Using combined PFGE and MLVA profiles as a molecular typing standard, we assessed the individual ability of WGS-based methods (core single nt variants (SNV) and k-mer analysis) or virulence gene profiling to differentiate sporadic cases from simultaneously occurring *E. coli* O157:H7 outbreak clusters.

### Methods

#### Molecular detection of *Escherichia coli* O157:H7 outbreaks in Alberta

Following established protocols from frontline microbiology diagnostic laboratories for enteric bacteria isolation, all presumptive *E. coli* O157:H7 isolates are forwarded to the Alberta Provincial Laboratory for Public Health (ProvLab) for serotype confirmation and molecular typing. Routinely, all *E. coli* O157:H7 isolates are subjected to PFGE and MLVA using standardised PulseNet protocols ([www.pulsenetinternational.org](http://www.pulsenetinternational.org)). For *E. coli*, XbaI endonuclease is the primary restriction enzyme used for chromosomal DNA digestion and is followed by secondary enzyme digestion with *Bln*I. Images (tagged image file format) of the PFGE profiles for all isolates are uploaded to the PNC Public Health Agency of Canada National Microbiology Laboratory in Winnipeg, Manitoba (PHAC-NML) secure national database for national pattern designation. Participating PNC public health laboratories across Canada are alerted of clusters (n=2 indistinguishable patterns) through the PNC web discussion board. In Alberta, the identification of STEC PFGE clusters triggers a public health investigation, involving the local public health authority and the Alberta ProvLab. MLVA analysis is performed at PHAC-NML on all *E. coli* O157:H7-confirmed isolates. A PFGE and MLVA cluster is defined as isolates with indistinguishable PFGE and MLVA patterns.

#### Outbreak identification

In Alberta, the Medical Officer of Health is notified of each case of *E. coli* O157:H7 in the province, which prompts an investigation into the case by an Alberta...
Health Services Environmental Health Officer who uses a standard questionnaire to identify potential sources of exposure. An outbreak investigation is initiated when a group of E. coli O157:H7 cases is identified with a common source or the aforementioned molecular typing criteria are met. ProvLab and public health officials are updated on the epidemiological and laboratory investigations through teleconferences, an online portal, and the distribution of line lists.

**Confirmation of Escherichia coli O157:H7 identification**

Tests performed on all isolates to identify and confirm the presumptive E. coli O157:H7 or to characterize the isolates further included: mauve colour colony on BBL CHROMagar O157 (Becton, Dickinson and Co., Mississauga, ON, Canada), haemolysis on blood agar, biochemical profiles using API20E (bioMérieux, Marcy l’Étoile, France), beta-glucuronidase expression, motility, sorbitol fermentation, citrate utilisation, malonate fermentation, tryptophanase production, serotyping using O157 direct antibody agglutination (BD Difco, Burlington, ON, Canada), and H7 antiserum (BD Difco) by tube flocculation.

**Genomic DNA isolation for virulence factor detection**

A single colony pick of E. coli O157:H7 grown overnight on 5% sheep blood agar (Dalynn Biologicals, Calgary, AB, Canada) at 35 to 37°C in ambient air was used for DNA extraction using the MagaZorb DNA Mini-Prep Kit (Promega Corporation, Madison, WI, US) on the KingFisher Magnetic Particle Processor (Thermo Scientific, Mississauga, ON, Canada). Purified DNA was used as template for the Stx typing and virulence gene dynamic array.

**Shiga toxin typing**

Detection of Stx genes, stx₁ and stx₂ was determined using a real-time multiplex polymerase chain reaction (PCR) assay consisting of two separate reactions run on an ABI Prism 7500FAST Sequence Detection System (Life Technologies, Inc., Burlington, ON, Canada) as previously described [9,10]. Conventional PCR was used to subtype stx₁ and stx₂ using primers from the World Health Organization Collaborating Centre for Reference and Research on Escherichia and Klebsiella [11].

**Virulence gene dynamic array**

The BioMark real-time PCR system (Fluidigm, San Francisco, CA, US) was used for real-time PCR
amplification of 49 genetic markers using 48.48 dynamic arrays. Amplifications were performed using the dyes, 6-carboxyfluorescein (FAM) - and 6-carboxy-2’, 4’, 5’, 7’, 7’-hexachlorofluorescein succinimidy-l ester (HEX)-labelled TaqMan probes as previously described [12]. All amplification assays included both positive and negative controls. Primers and probes used in this study have been described previously and target genetic markers derived from different plasmids and chromosomal loci [12-16]. Genetic markers included genes coding for the O157 antigen (rfbO157), the flagellar antigen H7 (flaH7), the long polar fimbriae (lpfA), the protein fimbriae of sorbitol-fermenting STEC O157 (spA), Stx 1 and 2 (stx1 and stx2), the pO157 markers (ehxA, katP, espP, toxB, etpB, ecfP), the heat-stable enterotoxin EAST1 (asta), adhesion factors (eae, eae-gamma, iha), effectors translocated by the type III secretion system derived from O1-122 (pagC, ent/espL2, nleB, nleE, efa1, efa2, Z4331), O1-50 (espK, espN, espO1-1, espX7), O1-71 (ecsi822, espM, nleF, nleA, nleH1-2), O1-37 (espX2), O1-44 (espV), O1-108 (espM2, espW), O1-153 (espY4-2), O1-174 (espX6), bacterial tel-lurite resistance (terE), urease (ureD), as well as open reading frames of the O1-57 (ecsi1763, Z2096, Z2098, Z2099, Z2121), clustered regularly interspersed short palindromic repeats (CRISPR) associated markers: SP.0157-A, SP.0157-B, and SP.0157-C, and the reference genetic marker for E. coli (wecA).

**Genome sequencing and assembly**
WGS of isolates was performed at the PHAC-NML Core Genomics facility. Sample libraries were prepared using Nextera XT library preparation kit (llumina, Inc., San Diego, CA, US). Sequencing was performed on the Illumina MiSeq platform with the MiSeq Reagent Kit V2 to achieve an average genome coverage greater than 50x for all isolates. Raw sequence reads are available under National Center for Biotechnology Information (NCBI) Bioproject PRJNA291542.

Sequencing reads were de novo assembled into contigs using SPAdes [17] and annotated with Prokka [18]. SPAdes-assembled contigs smaller than 1 kb were removed from the analysis.

**Core single nucleotide variant data preprocessing, quality control and data reduction**
All read data available for each genome were processed using the following steps: (i) FFastq files were converted to Sanger quality encoding Fastq format, (ii) all Fastq files for each isolate were concatenated into one Fastq file per isolate. The concatenated Fastq files were next subjected to a quality control step using a custom Perl script that trims the reads up to a maximum of 10 bases on either end of the reads if the average base call quality in that region was below 25. Next, reads less than 36 bp in length and reads with an average base quality call below 25 were discarded from the analysis. Following the quality control step, all data for each isolate were reduced to a maximum of 200 x coverage (estimated based on the total bp length of the E. coli O157:H7 strain Sakai genome, including a 92,721-bp plasmid, pO157) by random selection of reads.

**Core single nucleotide variant calling**
Core genome analysis was performed using the PHAC-NML bioinformatics custom Single Nt Variant Phylogenomic pipeline (SNVPhyl) [19] consisting of open-source software and custom Perl scripts. Briefly, sequencing reads were mapped against the complete reference genome, E. coli O157:H7 strain Sakai using SMALT v0.7.0.1.20 with a k-mer size of 13, a step size 6, and a minimum alignment fraction of 0.5. Variants were called using FreeBayes v0.9.8 [21] with variant reporting for all variants, no complex variants, minimum mapping quality of 30, minimum base quality of 30, minimum alternate fraction of variant bases in agreement of 75%, and minimum coverage of 20 reads at every position in the reference sequence.

**Core single nucleotide variant phylogeny**
For each of the variant call format files created by FreeBayes, complex variant calls were split into single variant calls to create new variant call format files. All variant calls were merged into a single alignment file. SAMTools v 0.1.18 [22] was used to investigate positions where not every genome had a variant call. Positions where no variant was called with SAMtools mpileup and with at least a minimum coverage of 20 were assigned the reference base in the alignment. Positions where a variant was called with SAMtools mpileup and with a minimum coverage of 20 were excluded from the alignment. Other positions were excluded. The alignment files were used to generate a phylogenetic tree with PhyML v3.0 [23] using a generalised time reversible (GTR) model and the best of both nearest neighbour interchange (NNI) and Subtree Pruning and Regrafting (SPR) tree topology searching strategies. Horizontally transferred elements arising from recombination events were identified using Progressive Mauve [24], PHAge Search Tool (PHAST) [25] and Island Viewer [26]. These coordinates were then masked from the SNV PHYLogenomics (SNVPhyl) analysis as were repetitive regions identified using the nucmer programme in the Mummer sequence alignment package.

**Core single nucleotide variant minimum spanning trees**
The high-quality (hq) core SNV for each isolate, identified in the SNVPhyl pipeline and used to generate the core SNV phylogenies were also visualised using minimum spanning trees (MST) generated with the open source Phyloviz goeBurst algorithm [27]. Each unique set of hq core SNVs and corresponding isolates were assigned a unique identifier or sequence type, and a table of core SNV positions and the unique sequence types were entered into the Phyloviz goeBURST algorithm. Additional metadata, including outbreak event, MLVA and PFGE patterns were annotated onto the MST.
Provincial Laboratory for Public Health, 14 July–17 September 2014 (n=140 isolates)

Outbreaks A, B (including cluster-α), and C are labelled by a specified colour. Sporadic and E. coli O157:H7 isolates submitted to the Alberta Provincial Laboratory for Public Health, 14 July–17 September 2014 (n=140 isolates)

Outbreaks A, B (including cluster-α), and C are labelled by a specified colour. Sporadic and E. coli O157:H7 isolates submitted to the Alberta Provincial Laboratory for Public Health, 14 July–17 September 2014 (n=140 isolates)

Figure 4
k-mer phylogeny, represented as a minimum spanning tree, of Escherichia coli O157:H7 isolates submitted to the Alberta Provincial Laboratory for Public Health, 14 July–17 September 2014 (n=140 isolates)

Outbreaks A, B (including cluster-α), and C are labelled by a specified colour. Sporadic and E. coli O157:H7 str. Sakai branches are in black. 14–6547 and 14–6537 (in black) are a cluster of two isolates unrelated to other isolates with exposure to ground beef.
**k-mer clustering**
The frequencies of all nt sequences of predefined lengths (k-mers) in the entire genome of each isolate was compared with the frequency of k-mers in all other isolates to determine a k-mer phylogeny tree. The alignment-free feature frequency profile (FFP) method [28] was used at a k-mer length of 25 nt and the SPAdes assembled contigs for each isolate. The optimal k-mer length was determined using the ffpvoacb utility part of the FFP package and the ffpreprof –e32 to determine the upper bound and the ffpreprof –f 2 to determine the lower bound. Phylogenetic trees were inferred from the resulting divergence distance matrix using the neighbour-joining method installed in the Phylip [29] package.

**Results**

**Description of the outbreaks**
From 14 July to 17 September 2014, 149 clinical isolates of *E. coli* O157:H7 were received by the Alberta ProvLab for molecular typing (Figure 1). Two outbreaks (designated as A and B) were identified in Alberta during this time period along with multiple sporadic isolates that were unrelated to either outbreak based on definitions involving epidemiological and PFGE/MLVA profiles. For comparison to an outbreak not temporally associated with outbreak A and B, a 2012 outbreak (outbreak C) was also included in this study.

The first outbreak (outbreak A), involved ten patients (four of whom developed HUS) with isolates collected from 14 to 22 July. Epidemiological investigations revealed a common exposure of visiting the same summer fair. There was one unique PFGE pattern combination and a total of three MLVA patterns among outbreak A isolates (Table and Figure 2). No food or environmental isolates were recovered in regards to this outbreak.

The second outbreak (outbreak B) was of particular interest because it was the largest in Alberta since PNC’s inception in the year 2000. In 2014, as of 17 September, 182 clinical isolates of *E. coli* O157:H7 were collected and referred to the Alberta ProvLab, whereas the mean annual case number for the previous five years was 87.6 (95% confidence interval: +/− 10.99). One hundred and eleven of these cases were associated with outbreak B during this time period, in which five developed HUS. Epidemiological investigations revealed a common exposure to contaminated pork products that were produced and distributed in Alberta [30]. This finding was confirmed by matching PFGE and MLVA profiles in human and food/environmental outbreak isolates (five food and one environmental isolate were received by Alberta ProvLab in regards to this outbreak).

PulseNet Canada routine surveillance using PFGE analysis revealed the presence of two concurrent clusters with closely related PFGE patterns (ECXAI.0023, ECBNI.0430 and ECXAI.0634, ECBNI.430) (Figure 2). As more isolates were received at the ProvLab, PFGE and MLVA subtyping revealed that the outbreak consisted of several isolates with variant, yet closely related PFGE and MLVA patterns (Figure 2).

The two most common PFGE/MLVA profiles were ECXAI.0023, ECBNI.0430/13_9_18_2_8_6_11_8 with 70 isolates and ECXAI.0634, ECBNI.430/13_9_19_2_8_6_11_8 with 14 isolates (Table). The most predominant difference in MLVA profiles was at the third locus, which varied from 17 to 20 repeats (Table). Further complicating the analysis were isolates collected from household members that were found to possess variant PFGE/MLVA profiles. For instance, isolates from one family with a HUS case all had the PFGE pattern ECXAI.0230, ECBNI.0430, however two MLVA patterns were observed, 13_9_19_2_8_6_11_8 (n=1) and 13_9_20_2_8_6_11_8 (n=2, including the HUS case). Another family with two infected members had their isolates collected on the same day with different PFGE/MLVA profiles (ECXAI.0230, ECBNI.0430/13_9_18_2_8_6_11_8 and ECXAI.0634, ECBNI.0430/13_9_19_2_8_6_11_8). No temporal associations were observed with the variant PFGE/MLVA profiles. There were also no distinct PFGE or MLVA patterns among HUS cases, which had the PFGE patterns ECXAI.0230 or ECXAI.0634, ECBNI.0430, and MLVA profile 13_9_19_2_8_6_11_8 except for one HUS isolate with the MLVA profile 13_9_20_2_8_6_11_8. Overall, 23 different PFGE and MLVA profiles and five clusters (two or more isolates with the same profile) were identified in outbreak B.

Outbreak B included a cluster of two isolates that was designated cluster-α. One isolate, 14–7110, recovered from a swab of a food tray at a distributor under investigation for a possible link with outbreak B was PFGE typed with an indistinguishable *Bln*I DNA restriction pattern as the primary outbreak B pattern (ECBNI.0430), but a variant, though closely related *Xba*I pattern (ECXAI.2098), and a variant MLVA pattern from the other outbreak B isolates (15_9_18_3_8_7_6_8) (Table and Figure 2). This isolate (14–7110) was collected one month after an isolate (14–5378) was recovered from an individual with an indistinguishable PFGE/MLVA profile. An epidemiological linkage between cluster-α and outbreak B was demonstrated in part by the recovery of an isolate (14–5369) with the most predominant difference in MLVA profiles was at the third locus, which varied from 17 to 20 repeats (Table). Further complicating the analysis were isolates collected from household members that were found to possess variant PFGE/MLVA profiles. For instance, isolates from one family with a HUS case all had the PFGE pattern ECXAI.0230, ECBNI.0430, however two MLVA patterns were observed, 13_9_19_2_8_6_11_8 (n=1) and 13_9_20_2_8_6_11_8 (n=2, including the HUS case). Another family with two infected members had their isolates collected on the same day with different PFGE/MLVA profiles (ECXAI.0230, ECBNI.0430/13_9_18_2_8_6_11_8 and ECXAI.0634, ECBNI.0430/13_9_19_2_8_6_11_8). No temporal associations were observed with the variant PFGE/MLVA profiles. There were also no distinct PFGE or MLVA patterns among HUS cases, which had the PFGE patterns ECXAI.0230 or ECXAI.0634, ECBNI.0430, and MLVA profile 13_9_19_2_8_6_11_8 except for one HUS isolate with the MLVA profile 13_9_20_2_8_6_11_8. Overall, 23 different PFGE and MLVA profiles and five clusters (two or more isolates with the same profile) were identified in outbreak B.

In 2012, there was a beef-associated *E. coli* O157:H7 outbreak with great social and economic impact involving 18 human-clinical cases across Canada and no cases of HUS (outbreak C) [31]. All human-clinical and beef isolates in this outbreak had one indistinguishable PFGE/MLVA profile (Table and Figure 2).
### Table

Number of samples for each pulsed-field gel electrophoresis (PFGE) and multilocus variable-number tandem repeat analysis (MLVA) profile found in outbreaks A, B, (14 July–17 September 2014) and C (2012), Alberta, Canada (n=136)

<table>
<thead>
<tr>
<th>Outbreak/ Type of isolate</th>
<th>Xba&lt;sup&gt;*&lt;/sup&gt;</th>
<th>Bln&lt;sup&gt;*&lt;/sup&gt;</th>
<th>MLVA</th>
<th>N</th>
<th>Location on core SNV tree&lt;sup&gt;a&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>A</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clinical</td>
<td>1897</td>
<td>126</td>
<td>8...</td>
<td>6</td>
<td>Main outbreak A cluster</td>
</tr>
<tr>
<td></td>
<td>1897</td>
<td>126</td>
<td>8...</td>
<td>2</td>
<td>3 SNV from main outbreak A cluster</td>
</tr>
<tr>
<td></td>
<td>1897</td>
<td>126</td>
<td>8...</td>
<td>2</td>
<td>1–2 SNV from main outbreak A cluster</td>
</tr>
<tr>
<td><strong>B</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clinical</td>
<td>23</td>
<td>430</td>
<td>13...</td>
<td>70</td>
<td>Cluster 74 (n=59), cluster 23 (n=1), 1–2 SNV from cluster 74: 14–6228, 14–5502, 14–5769, 14–5384, 14–5503, 14–5376, 14–5392, 14–7108, 14–5643, 14–6099</td>
</tr>
<tr>
<td></td>
<td>634</td>
<td>430</td>
<td>13...</td>
<td>14</td>
<td>Cluster 23 (n=13), 1 SNV from cluster 23: 14-6232</td>
</tr>
<tr>
<td></td>
<td>634</td>
<td>430</td>
<td>13...</td>
<td>3</td>
<td>Cluster 74 (n=2), cluster 23 (n=1)</td>
</tr>
<tr>
<td></td>
<td>634</td>
<td>1305</td>
<td>13...</td>
<td>3</td>
<td>Cluster 23</td>
</tr>
<tr>
<td></td>
<td>634</td>
<td>430</td>
<td>13...</td>
<td>3</td>
<td>Cluster 23</td>
</tr>
<tr>
<td></td>
<td>23</td>
<td>75</td>
<td>13...</td>
<td>1</td>
<td>1 SNV from cluster 74: 14–5768</td>
</tr>
<tr>
<td></td>
<td>23</td>
<td>310</td>
<td>13...</td>
<td>1</td>
<td>1 SNV from cluster 74: 14–6372</td>
</tr>
<tr>
<td></td>
<td>23</td>
<td>430</td>
<td>13...</td>
<td>1</td>
<td>Cluster 74</td>
</tr>
<tr>
<td></td>
<td>23</td>
<td>430</td>
<td>13...</td>
<td>1</td>
<td>Cluster 74</td>
</tr>
<tr>
<td></td>
<td>23</td>
<td>430</td>
<td>14...</td>
<td>1</td>
<td>Cluster 74</td>
</tr>
<tr>
<td></td>
<td>23</td>
<td>592</td>
<td>13...</td>
<td>1</td>
<td>Cluster 74</td>
</tr>
<tr>
<td></td>
<td>23</td>
<td>1302</td>
<td>13...</td>
<td>1</td>
<td>1 SNV from cluster 74: 14–5377</td>
</tr>
<tr>
<td></td>
<td>597</td>
<td>430</td>
<td>13...</td>
<td>1</td>
<td>Cluster 74</td>
</tr>
<tr>
<td></td>
<td>634</td>
<td>430</td>
<td>13...</td>
<td>1</td>
<td>Cluster 74</td>
</tr>
<tr>
<td></td>
<td>634</td>
<td>430</td>
<td>13...</td>
<td>1</td>
<td>1 SNV from cluster 23: 14–5398</td>
</tr>
<tr>
<td></td>
<td>634</td>
<td>430</td>
<td>13...</td>
<td>1</td>
<td>Cluster 23</td>
</tr>
<tr>
<td></td>
<td>2113</td>
<td>1300</td>
<td>13...</td>
<td>1</td>
<td>Cluster 74</td>
</tr>
<tr>
<td></td>
<td>2592</td>
<td>735</td>
<td>13...</td>
<td>1</td>
<td>Cluster 74</td>
</tr>
<tr>
<td></td>
<td>3106</td>
<td>406</td>
<td>13...</td>
<td>1</td>
<td>1 SNV from cluster 74: 14–5370</td>
</tr>
<tr>
<td></td>
<td>3112</td>
<td>430</td>
<td>13...</td>
<td>1</td>
<td>Cluster 74</td>
</tr>
<tr>
<td></td>
<td>3112</td>
<td>430</td>
<td>13...</td>
<td>1</td>
<td>1 SNV from cluster 23: 14–7107</td>
</tr>
<tr>
<td></td>
<td>2098</td>
<td>430</td>
<td>15...</td>
<td>1</td>
<td>Cluster-α</td>
</tr>
<tr>
<td>Pork</td>
<td>23</td>
<td>430</td>
<td>13...</td>
<td>4</td>
<td>Cluster 74 (n=3), 1 SNV from cluster 74: 14–6364</td>
</tr>
<tr>
<td></td>
<td>634</td>
<td>430</td>
<td>13...</td>
<td>1</td>
<td>Cluster 23</td>
</tr>
<tr>
<td>Environmental</td>
<td>2098</td>
<td>430</td>
<td>15...</td>
<td>1</td>
<td>Cluster-α</td>
</tr>
<tr>
<td><strong>C</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clinical</td>
<td>1</td>
<td>12</td>
<td>8...</td>
<td>7</td>
<td>Outbreak C cluster</td>
</tr>
<tr>
<td>Beef</td>
<td>1</td>
<td>12</td>
<td>8...</td>
<td>2</td>
<td>Outbreak C cluster</td>
</tr>
</tbody>
</table>

MLVA: multilocus variable-number tandem repeat analysis; PFGE: pulsed-field gel electrophoresis; SNV: single nt variant.

The MLVA and PFGE profiles for the food isolates (pork or beef) and the one environmental isolate from outbreak B (14–7110) are shown separately from the clinical clusters.

<sup>a</sup> PFGE pattern designation without the preceding zero (s).

<sup>b</sup> Location of the PFGE/MVA profiles in the core SNV minimum spanning tree (Figure 3). Of the 111 clinical, five food and one environmental outbreak B-associated isolates sequenced, there were two major clusters in the core SNV minimum spanning tree, one with 74 isolates (cluster 74) and the other with 23 isolates (cluster 23), whereby no core SNV differences between the isolates occurred within the clusters. Outbreak B isolates that were not part of a core SNV cluster, but differed by 1–2 core SNVs from a core SNV cluster are designated as 1 or 2 SNV from cluster 74 or 1 SNV from cluster 23 with the isolate number(s). Cluster 74 included three pork isolates (14–6540, 14–6549 and 14–7569) and 71 clinical isolates with the following laboratory accession numbers: 14–5190 to 91, 14–5196, 14–5369, 14–5371, 14–5373, 14–5376 to 78, 14–5379 to 85, 14–5385 to 87, 14–5393 to 5399, 14–5501 to 03, 14–5693, 14–5695 to 00, 14–5696, 14–5697 to 07, 14–5699 to 11, 14–5613, 14–5615 and 16, 14–5619 to 21, 14–5625 to 26, 14–5761 to 67, 14–5866 to 88, 14–5891, 14–6200, 14–6340 to 71, 14–6373, 14–6541, 14–6543–45, 14–6548, 14–7104, 14–7109. Cluster 23 included one pork isolate (14–6539) and 22 clinical isolates: 14–5192, 14–5198, 14–5200, 14–5390, 14–5491, 14–5501, 14–5601 to 02, 14–5608, 14–5612, 14–5617, 14–5622, 14–5770, 14–5890, 14–5892, 14–6101, 14–6229 to 30, 14–6365, 14–6368, 14–7106, 14–7813.
Whole genome sequencing
The genetic relationships among outbreak-associated isolates were also determined by WGS using the PHAC-NML bioinformatics SNVPhyl pipeline and by k-mer tree analysis. All aforementioned clinical, food, and environmental isolates were sequenced for each outbreak. For outbreak C, isolates from eight respective human cases that had occurred in Alberta and two food isolates were available for sequencing. One clinical isolate was excluded because it did not meet the sequence quality threshold. The sporadic isolates (no epidemiological or PFGE/MLVA link to an outbreak) that were sequenced included a cluster of two isolates unrelated to any outbreak with suspected exposure to ground beef (14–6537 and 6547), one isolate (14–5400) that came from an individual who worked with pork and one isolate (14–5618) that was randomly chosen.

Core single nucleotide variant phylogenetic analysis
The interrogation of hq core SNVs revealed that clinical human isolates within an outbreak event varied by 0–5 SNVs from one another other (0–5 SNVs for outbreak A and 0–5 SNVs for outbreak B) (Figure 3). Food/environmental isolates from outbreak B also clustered within 0–5 SNVs from the clinical human isolates. Outbreak A and outbreak B isolates clustered into distinct and well-defined branches, separated by a distance of 231–257 SNVs (Figure 3). Both outbreaks also clustered away from the sporadic isolates and the reference strain, E. coli O157:H7 str. Sakai.

Within outbreak B, the major PFGE/MLVA clusters ECXAI.0634, ECBNI.0430/13_9_19_2_8_6_11_8 and ECXAI.0023, ECBNI.0430/13_9_18_2_8_6_11_8 were separated by 2–4 SNV (Table and Figure 3). The HUS-associated isolates from outbreak B clustered with the other outbreak isolates with four in cluster 74 and one in cluster 23 (Figure 3). Upon further examination of the core SNV phylogenetic trees, the two isolates composing cluster-α were separated by 22–27 hq core SNVs from all other outbreak B isolates and were distanced by only one core SNV from each other (Figure 3). The isolate that was not part of cluster-α, but isolated from a diner at the same restaurant (14–5369) was found in the cluster of 74 isolates.

Performing core SNP analysis on the seven Alberta clinical human isolates involved in outbreak C (the 2012 beef-related outbreak) and two food isolates, revealed no core SNP differences between these nine outbreak-associated isolates. Moreover, this method differentiated the branch corresponding to the outbreak C isolates from the outbreak B branch by a distance of 157–161 core SNVs and outbreak A by 74–77 core SNVs.

Several sporadic isolates were identified concurrent with outbreaks A and B, and possessed variant PFGE/MLVA profiles. These isolates have a minimum genetic distance of 57 hq core SNVs from all three outbreak clusters. Notably, one sporadic case (14–5400) recovered from an individual who worked directly with pork had a similar, but distinguishable BlnI pattern to outbreak A and a distinct MLVA profile (ECXAI.3108, ECBNI.0181/9_10_12_7_7_6_3_6) (denoted as Sporadic I in Figure 2). Core SNP phylogenetic analysis demonstrated that this isolate was distant from outbreak B isolates by a high number of core SNVs, 214–237.

k-mer clustering
The k-mer method also delineated outbreak A from outbreak B with each group of outbreak isolates located on distantly related nodes (Figure 4). There were isolates found within the outbreak B node (orange in Figure 4) with greater horizontal branch distances than the average branch lengths for outbreak B isolates, demonstrating additional resolution of certain isolates from other outbreak B isolates. These more distant isolates included ones with the most and least frequently observed MLVA and PFGE profiles. Cluster-α was also distinguished from the main outbreak B branch and was identified in a separate node, but originated from the same site as the outbreak B branch and thus, share a most recent common ancestor (Figure 4). In concordance with core SNP analysis, the isolate that was not part of cluster-α, but isolated from a diner at the same restaurant (14–5369) was not differentiated from other outbreak B isolates by k-mer clustering. k-mer analysis also did not indicate a close relationship between sporadic isolates and outbreak-associated isolates.

Shiga toxin genes typing
The presence of stx1 and stx2 was determined for all isolates received during the sampling period. For stx1 and stx2 subtyping, 36 isolates were selected from the 2014 sampling period comprising of outbreak-related food isolates (n = 3), HUS-associated isolates (n = 2 for outbreak A and n = 5 for outbreak B), all isolates from household members of HUS cases (based on address) (n = 4, all outbreak B), both the environmental and clinical isolate from cluster-α, a representative isolate for each outbreak B variant PFGE and MLVA profile not covered by other selection criteria (n = 19), and one isolate (14–6547) from the sporadic cluster of two individuals. In addition, all clinical and food isolates from outbreak C were subtyped (n = 10). All outbreak B isolates (n = 3) were stx2 negative and stx1 positive (subtype stx2a). Most outbreak B isolates (27 of 31), including HUS-associated isolates were positive for both stx genes and subtyped as stx1a and stx2a. The clinical human isolate from cluster-α (14–5376) tested positive for stx1a, stx2a and stx2c, and the environmental isolate in this cluster (14–7110) was stx1a and stx2a positive, but no subtype was identified for stx2c. The other variant stx2 subtypes in outbreak B included two clinical isolates, one stx2/a/stx2c untypeable and one stx1 negative /stx2c untypeable. All outbreak C isolates and isolate 14–6547 from the sporadic cluster were stx1a and stx2a positive.

Virulence gene profiling
Virulence gene profiling was performed on all isolates (n = 155 received by ProvLab for molecular typing from
14 July to 17 September 2014. Outbreak A, outbreak B and sporadic isolates were indistinguishable based on gene profiling of 49 STEC virulence genes. All isolates were negative for the pilin subunit gene found in sorbitol fermenting STEC, sfpA and all but six isolates were positive for all other genes tested. Two outbreak B isolates, 14–6543 and 14–5377, tested negative for the putative proteins ZZ2096 and ZZ2098. In the k-mer analysis result, these isolates clustered within outbreak B, but both were distinguishable from each other and all other outbreak B isolates (Figure 4). A single outbreak B isolate, 14–5380, tested negative for the secreted effector proteins, espK, espN, espX7, and esp01–1 and was found to be related to outbreak B, but distinguishable from all other outbreak B isolates by k-mer analysis (Figure 4). All of the outbreak B isolates with genes not detected on the virulence gene array were indistinguishable from other outbreak B isolates using SNV analysis. The other three isolates that typed negative for virulence genes were sporadic isolates with the first testing negative for espX7, the second negative for espP, and the third testing negative for efa1, efa2, ent, nleB, nleE, pagC, and Z4331. None of the HUS cases were negative for any virulence genes other than sfpA.

**Discussion**

Technologies such as high throughput screening of virulence genes [32,33] and WGS have the potential to be used for early outbreak detection and characterisation. To be used as such, these methods must be compared with current international standards of bacterial typing. For *E. coli* O157:H7, the currently validated and widely employed standard typing methods are MLVA and PFGE. Furthermore, a bacterial typing method must have the following characteristics: accuracy, inter and intra-laboratory reproducibility (including with multiple passaging of isolates), high discriminatory power, concordance with epidemiological data, rapid and ease-of-use, cost effectiveness, and amenability to computerised analysis [34]. This work demonstrates that for typing of *E. coli* O157:H7, WGS is a suitable typing method already meeting many of these criteria, but not virulence gene profiling involving the genetic targets used in this study.

No virulence factors tested in this work could reliably distinguish between outbreak and non-outbreak strains. The high prevalence of the genetic markers used in this study in *E. coli* O157:H7 (reference 13 and this study), prevents the use of any assay that detects the presence or absence of these genes for the purpose of accurate discrimination between isolates. Shiga toxin subtyping was able to differentiate isolates only in cases where the subtypes differed from the most prevalent subtypes (stx1 and stx2) a result similarly observed in an analysis of Alberta E. coli O157:H7 isolates collected from 2004 to 2012 [15].

Phylogenetic analysis using whole genome or core SNVs derived by a variety of methods is the most frequently published method to determine relatedness between isolates of *E. coli* or other Enterobacteriaceae such as *Salmonella* species [33-38]. To the best of our knowledge, this is the first example of a real-time, large-scale study comparing virulence gene profiling, hq core SNV or k-mer analysis to MLVA and PFGE profiling.

Using different bioinformatics pipelines and groups of *E. coli* isolates, four previously published studies have demonstrated the ability of WGS to discriminate between isolates of *E. coli* [35-38]. Core SNV phylogeny cannot only differentiate STEC isolates from other STEC isolates, but also uropathogenic *E. coli* from one another [35]. Two other studies have used MLVA as the ‘standard’ reference method and demonstrated core SNV phylogenies to be equivalent or better than MLVA at identifying outbreaks [36,37]. One of these two studies used *E. coli* O157:H7 isolates from the United Kingdom and identified outbreaks using core SNV analysis that were missed by epidemiological investigations and MLVA analysis [37]. Another study analysing isolates from a beef-associated outbreak of *E. coli* O157:H7 in Denmark demonstrated that core SNV phylogenetic analysis and nt-distance based trees (built using 17-base k-mers) methods were each capable of differentiating between outbreak isolates and concurrently occurring, non-outbreak sporadic *E. coli* O157:H7 and non-O157 STEC isolates. In the Danish study, PFGE was also performed on selected isolates, but its utility for outbreak detection was not compared with the two types of WGS approaches [38]. The study herein adds to this knowledge by demonstrating that core SNV or k-mer phylogenies alone are concordant with combined PFGE and MLVA data when used to differentiate between concurrently occurring outbreak and non-outbreak isolates of *E. coli* O157:H7.

k-mer analysis showed concordance with PFGE and MLVA profiling while advantageously revealing increased discriminatory power when compared with PFGE and MLVA profiling or core SNV. Feature frequency profiling generates k-mers-based profiles using the entire genomic sequence of test organisms to determine relatedness, whereas core SNV analysis uses only the ‘conserved’ portion of the genome; therefore one may hypothesise that k-mer analysis would provide additional discriminatory power. With this increased discriminatory power, there is the concomitant risk that k-mer analysis may be too discriminatory by including not only ‘core’ conserved features, but also genetic elements resulting from horizontal gene transfer. These elements can be easily lost and gained as isolates undergo natural and passaging in the laboratory. Therefore, the inclusion of these elements in the analysis has the potential to mask core phylogenetic inference, but no evidence of this was observed in our study.

Cluster-a consisted of two isolates that differed from all other outbreak B isolates by at least 22 SNV, but differed from each other by only one core SNV. This cluster was included in outbreak B based on PFGE and epidemiological links (an outbreak B implicated restaurant),
but was differentiated from other outbreak B isolates by MLVA, k-mers and core SNV analysis. Furthermore, the stx₂ subtype for these two isolates differed from each other and the other subtyped outbreak B isolates. This cluster likely represents different strains contributing to the outbreak. Similar observations have been made previously in our laboratory as well as by Gilmour et al. [39]. These observations emphasise the need to consider careful correlation of epidemiological together with molecular subtyping data.

The core SNV genetic distances observed between cluster-α and outbreak B exemplifies the utility of predefining a SNV threshold for isolate relatedness. Cluster-α differed from outbreak B by 22–27 SNVs whereas outbreak-unrelated, sporadic isolates differed by ≥ 58 SNVs. Despite using different SNV calling and phylogeny methods, other studies have also identified ≤ 5 SNVs as a potential threshold for genetic relatedness among E. coli [36,37]. In one study of extended-spectrum beta-lactamase-E. coli outbreak-associated isolates, 0 SNVs were identified between outbreak-associated isolates [35]. It is more difficult to assign a numerical ‘threshold’ or pattern for relatedness with k-mer analysis, which produces trees based on a distance matrix generated from k-mer profiles and their presence or absence. Unlike MLVA and/or PFGE using PulseNet standardised methods, both k-mer and core SNV methods do not generate a numerical ‘barcode’ for the organisms, which makes it difficult to compare isolates from different laboratories. Therefore, a large shared and curated database will likely be required to generate cluster identifiers that could act as the organism’s ‘barcode’.

Currently, PulseNet networks worldwide conduct timely surveillance and facilitate outbreak detection. For WGS to be used in this manner, standardised protocols, quality validation metrics and a robust method to determine isolate relatedness will need to be established. Published reports use different sequencing platforms and bioinformatics pipelines to assemble and analyse WGS data, prohibiting direct comparisons between studies. Head-to-head comparison of these different methods will help determine the appropriate standard(s). The inter-laboratory comparison of WGS data also requires communal database of isolates, which requires adequate computing infrastructure and secure electronic networks capable of transmitting large datasets. Other considerations encompass the ethical, legal or political barriers to sharing complete genomics data between various health authorities.

In conclusion, WGS holds significant potential to replace current gold-standard typing methods such as PFGE for the routine surveillance and detection of enteric outbreaks. This shift will be driven by the advantages offered by WGS such as increased discriminatory power and genetic resolution. However, before this technology can be widely implemented, certain barriers remain to be addressed such as initial capital expense, computing infrastructure, and validated automated, user-friendly WGS analysis software. Moreover, quality metrics and standardised protocols including standardised definitions for isolate relatedness are required before routine application of WGS within public health laboratories.

Acknowledgements

Provlab & PHAC-NML laboratory technicians who performed the PFGE, MLVA and laboratory work associated with WGS, Natalie Knox, Aaron Petkau, Philip Mabon, Morag Graham, and Gary van Domselaar at PHAC-NML who developed the bioinformatics pipeline. The Canada-Alberta Partners in Food Safety Outbreak Investigation Coordinating Committee who coordinated and carried out the outbreak investigations outside the laboratory. Alberta Health for patient demographic information.

Conflict of interest

None declared.

Authors’ contributions

All authors reviewed the manuscript drafts and approved the final version. BMB designed the study, performed Shiga toxin subtyping and genomic DNA isolation for virulence gene profiling, analysed the data, prepared the figures and wrote the manuscript. TP and CB performed comparative genomics, the phylogenetic and minimum spanning tree assembly. CB contributed writing to sections of the manuscript. SD and PF designed the virulence gene profile and revised the manuscript. CN and LT performed PFGE pattern designation and MLVA analysis. LH reviewed the epidemiological data and revised the manuscript. ML reviewed and revised the manuscript. LC designed the study, coordinated data collection, analysed the data and wrote the manuscript.

References


www.eurosurveillance.org


