



THE UNIVERSITY *of* EDINBURGH

This thesis has been submitted in fulfilment of the requirements for a postgraduate degree (e. g. PhD, MPhil, DClinPsychol) at the University of Edinburgh. Please note the following terms and conditions of use:

- This work is protected by copyright and other intellectual property rights, which are retained by the thesis author, unless otherwise stated.
- A copy can be downloaded for personal non-commercial research or study, without prior permission or charge.
- This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the author.
- The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author.
- When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given.

The use of Oxford Nanopore technologies to investigate outbreaks of Shiga toxin-producing *Escherichia coli* O157:H7 in humans.

David Robert Greig

UK Health Security Agency

University of Edinburgh

Submitted for examination in the degree of
Doctor of Philosophy

2018 - 2025



UK Health
Security
Agency



Signed declaration:

I, David Robert Greig, confirm that the work presented in this thesis was composed by me and is my own. I confirm that the work has not been submitted for any other degree or professional qualification. Where information has been derived from other sources, I confirm that this has been indicated in the thesis. The included publications in section 8.3 are my own work.

David Robert Greig, October 2025.

Table of contents:

Signed declaration:	2
Table of contents:	3
List of figures:	9
List of tables:	13
Acknowledgements:	15
Layman's summary:	16
Thesis abstract:	17
Chapter 1: Introduction.	20
1.1 Pathogenic <i>Escherichia coli</i> .	21
1.2 Shiga toxin-producing <i>Escherichia coli</i> (STEC).	23
1.3 Shiga toxin-producing <i>Escherichia coli</i> (STEC) O157:H7.	26
1.4 Genome evolution of STEC O157:H7.	29
1.5 Detection and typing of STEC O157:H7.	34
1.6 Population structure of STEC O157:H7 in the UK.	36
1.7 Epidemiology of STEC O157:H7 in the UK.	37
1.8 Bacteriophages and prophages associated with STEC O157:H7.	39
1.9 Shiga-toxin encoding bacteriophages.	43
1.10 STEC O157:H7 genome plasticity.	45
1.11 Impact of whole genome sequencing on UK national surveillance of STEC O157:H7.	46
1.12 Novel sequencing technologies.	48
1.13 PhD aims.	52
Chapter 2: Integration of long and short-read sequencing data for the analysis of STEC O157:H7 genomes.	53
2.1 Chapter 2 background.	54
2.2 Chapter 2 methodology.	56
2.2.1 Assembly-dependent methodology.	56
2.2.1.1 Bacterial culture.	56
2.2.2.2 DNA extraction and library preparation for Illumina platforms.	56
2.2.2.3 DNA extraction and library preparation for Nanopore platforms.	56
2.2.2.4 Bioinformatics processing including <i>de novo</i> genome assembly.	57
2.2.2.5 Assembly correction workflows.	57
2.2.2.6 Training an STEC-specific Medaka assembly correction model.	60
2.2.2.7 Comparison of correction workflow outputs to truth set (Strain Sakai).	60
2.2.2.8 Workflow selection and subsequent optimisation for STEC genomes.	60
2.2.2.9 Genome annotation and visualisation.	61
2.2.3 Assembly-independent methodology.	62
2.2.3.1 Genomic DNA extraction, library preparation and Illumina sequencing.	62
2.2.3.2 Illumina sequencing data bioinformatics processing.	63

2.2.3.3 Genomic DNA extraction, library preparation and Nanopore sequencing.	63
2.2.3.4 Data processing and analysis of Nanopore sequence data.	64
2.2.3.5 Comparison of Illumina and Nanopore discrepant SNPs.	65
2.2.3.6 Generation of Phylogenetic trees.	65
2.3 Chapter 2 results.	67
2.3.1 Assembly-dependent results.	67
2.3.1.1 Initial draft assembly statistics and distance of true reference.	67
2.3.1.2 Initial workflow results and selection of highest scoring workflow.	67
2.3.1.3 Final workflow optimisation.	72
2.3.1.4 Final draft assemblies and description of remaining discrepancies.	73
2.3.1.3 Comparison of prediction genome annotations.	75
2.3.2 Assembly-independent results.	76
2.3.2.1 Optimisation of Nanopore variant calling.	76
2.3.2.2 Investigation of the discrepant variants identified between Illumina and Nanopore.	79
2.3.2.3 Phylogenetic analyses.	84
2.4 Chapter 2 discussion.	87
2.4.1 Assembly-based correction workflow optimisation.	87
2.4.2 Assembly-based correction discrepancies related to the STEC O157:H7 genome.	88
2.4.3 Assembly-independent methodology to determine STEC genome relatedness.	88
2.4.4 Assembly-independent methodology limitations and discrepancies.	89
2.4.5 Nanopore technological advancements.	90
2.5 Chapter 2 conclusions.	91
Chapter 3: Detection and characterisation of Shiga toxin-encoding prophages within highly pathogenic STEC O157:H7 UK sub-lineages.	93
3.1 Chapter 3 background.	94
3.2 Chapter 3 methods.	95
3.2.1 Strain selection and bacterial culture.	95
3.2.2 Illumina sequencing and data processing.	95
3.2.3 Nanopore sequencing and data processing.	96
3.2.4 <i>De novo</i> assembly, correction, and re-orientation.	97
3.2.5 Prophage detection and comparisons.	98
3.2.6 Mash and <i>stx</i>-encoding prophage phylogeny.	98
3.2.7 Visualisation tools and phylogenetic context.	98
3.3 Chapter 3 results.	100
3.3.1 Genomic features of Nanopore sequenced samples.	100
Finally, the <i>stx</i> subtypes associated with each strain were either <i>stx1a</i> and/or <i>stx2a</i> and/or <i>stx2c</i> (Table 3.1).	100
3.3.2 Comparison of the <i>stx1a</i>-encoding prophages.	103
3.3.3 Comparison of the <i>stx2c</i>-encoding prophages.	106

3.3.4 Comparison of the <i>stx2a</i> -encoding prophages.	108
3.3.5 Comparison of the prophages of the non-domestic strains.	109
3.4 Chapter 3 discussion.	111
3.4.1 <i>stx1a</i> -encoding prophages.	111
3.4.2 <i>stx2c</i> -encoding prophages.	111
3.4.3 <i>stx2a</i> -encoding prophages.	111
3.5 Chapter 3 conclusions.	113
Chapter 4: Revealing the intra-outbreak accessory genome variation of Shiga toxin-producing <i>E. coli</i> (STEC) O157:H7.	115
4.1 Chapter 4 background.	116
4.2 Chapter 4 methods.	118
4.2.1 Strain selection and preparation.	118
4.2.2 Illumina-based DNA extraction, library preparation, sequencing and data processing.	118
4.2.3 Nanopore-based DNA extraction, library preparation, sequencing and data processing.	119
4.2.4 <i>De novo</i> assembly, correction, reorientation, and annotation.	120
4.2.5 Prophage detection, excision, comparison and generation of neighbour-joining trees.	120
4.2.6 Variant calling and masking, SNP typing and generation of maximum-likelihood phylogenetic trees.	121
4.2.7 Detection and characterisation of chromosomal structural variation.	122
4.2.8 <i>In silico</i> plasmid typing and characterisation.	123
4.3 Chapter 4 results.	124
4.3.1 Outbreak associated with contaminated raw pet food.	124
4.3.1.1 Assembly metrics, variant calling and phylogeny.	124
4.3.1.2 Analysis of prophage content of the outbreak isolates and comparison.	126
4.3.1.3 Within-outbreak comparison of the prophage regions of the four outbreak isolates.	130
4.3.2 Outbreak associated with consumption of raw drinking milk.	132
4.3.2.1 Genome features of long-read assemblies.	132
4.3.2.2 Comparison of prophage and prophage-like content.	136
4.3.2.3 Analysis of Shiga toxin-encoding prophages.	138
4.3.2.4 Analysis of plasmid content.	140
4.3.2.5 Comparison of variant calling methods and phylogeny.	140
4.3.3 Outbreak associated with contaminated mud-based obstacle course.	143
4.3.3.1 Genome features of long-read draft assemblies.	143
4.3.3.2 Analysis of structural variation within outbreak samples.	145
4.3.3.3 Analysis of prophage and prophage-like content.	148
4.3.3.4 Analysis of Shiga toxin-encoding prophages.	151
4.3.3.5 Prophage-like regions and pathogenicity island variation.	154
4.3.3.6 Analysis of plasmid content.	155
4.3.3.7 Comparison of variant calling methods and phylogeny.	157

4.3.4 Outbreak associated with animal contact during a lambing event.	159
4.3.4.1 Genome features of long-read draft assemblies.	159
4.3.4.2 Analysis of structural variation within outbreak samples.	160
4.3.4.3 Analysis of prophage and prophage-like content.	161
4.3.4.4 Analysis of Shiga toxin-encoding prophages.	164
4.3.4.5 Analysis of plasmid content.	167
4.3.4.6 Comparison of variant calling methods and phylogeny.	167
4.4. Chapter 4 discussion.	170
4.4.1 Comparison of variant calling methods and phylogeny.	170
4.4.2 Outbreak associated with contaminated raw pet food.	171
4.4.3 Outbreak associated with consumption of raw drinking milk.	172
4.4.4 Outbreak associated with contaminated mud-based obstacle course.	173
4.4.5 Outbreak associated with animal contact during a lambing event.	175
4.5. Chapter 4 conclusions.	176
Chapter 5: The use of long-read sequencing to detect and characterise large chromosomal rearrangements in STEC O157:H7 genomes.	178
5.1. Chapter 5 background.	179
5.2. Chapter 5 methods.	180
5.2.1 Strain selection and culture.	180
5.2.2 Illumina-based DNA extraction, library preparation, sequencing and data processing.	180
5.2.3 Nanopore-based DNA extraction, library preparation, sequencing and data processing.	180
5.2.4 <i>De novo</i> assembly, correction, reorientation, and annotation.	181
5.2.5 Prophage detection, excision and comparison.	182
5.2.6 Detection and characterisation of chromosomal structural variation.	182
5.3. Chapter 5 results.	184
5.3.1 Chromosomal inversions.	184
5.3.3 Chromosomal inverted translocation.	188
5.3.3 Homologous prophage regions.	190
5.3.4 Compound prophages.	194
5.4. Chapter 5 discussion.	196
5.4.1 Prophage boundaries of LCRs.	196
5.4.2 Chromosomal isoforms.	198
5.4.3 Compound prophages.	200
5.4.4 Impact on <i>de novo</i> assembly and bioinformatics considerations.	201
5.5. Chapter 5 conclusions.	202
Chapter 6: Detection and characterisation of antimicrobial resistance determinants in STEC O157:H7 using long-read sequencing.	203

6.1. Chapter 6 background.	204
6.2. Chapter 6 Methods.	206
6.2.1 Strain selection and culture.	206
6.2.2 Illumina-based, DNA extraction, library preparation and sequencing.	206
6.2.3 Illumina-based bioinformatics processing.	206
6.2.4 Phylogenetic tree construction.	207
6.2.5 Nanopore-based, DNA extraction, library preparation and sequencing.	208
6.2.6 Nanopore-based bioinformatics processing.	208
6.2.7 <i>de novo</i> assembly and assembly correction.	209
6.2.8 <i>In silico</i> plasmid replicon typing, detection and characterisation of AMR determinants.	210
6.3. Chapter 6 results.	211
6.3.1 General presence of AMR determinates in STEC O157:H7.	211
6.3.2 Presence of AMR determinants by antibiotic class.	211
6.3.3 Phylogenetic analysis of AMR determinants and analysis of Nanopore-based data in sub-lineage I/II.	214
6.3.4 Phylogenetic analysis of AMR determinants and analysis of Nanopore-based data in sub-lineage Ib.	214
6.3.5 Phylogenetic analysis of AMR determinants and analysis of Nanopore-based data in sub-lineage IIc.	218
Black links denote genes of $\geq 90\%$ sequence similarity.	221
6.4. Chapter 6 discussion.	222
6.5. Chapter 6 conclusions.	224
Chapter 7: Discussion/Conclusions	225
7.1. Final discussion.	226
7.2. Limitations of analyses.	237
7.3. Proposed future works.	238
7.4. Final conclusions.	239
Reference List/Bibliography:	241
Chapter 8: Appendix	258
8.1 Data availability.	259
8.2 Software availability.	259
8.3 Published materials from PhD analyses.	259
Manuscript 1:	262
Manuscript 2:	274
Manuscript 3:	284
Manuscript 4:	296

Manuscript 5:	306
Manuscript 6:	316
Manuscript 7:	328

List of figures:

Chapter 1 figures:

Figure 1.1 – Schematic of the endocytosis and processing of the Shiga toxin and apoptotic pathway induction Adapted from Johannes and Römer 2010 [28].

Figure 1.2 – The first stepwise evolutionary model of STEC O157:H7 proposed by Feng *et al* 1998 [76].

Figure 1.3 – Maximum-likelihood phylogeny showing the grouping of STEC with similar genome features within the evolutionary model. From Kyle *et al* 2012 [79].

Figure 1.4 – Bacteriophage replication showing both lytic and lysogenic cycles in this case Phage- λ (from Campbell 2003 [108]).

Figure 1.5 – Chromosome of STEC O157:H7 strain Sakai with inner ring showing black blocks which equate to prophages and their relative positions on the chromosome (Hayashi *et al* 2001 [75]).

Figure 1.6 – Two genome assemblies of the same *E. coli* isolate. Top - assembled using short Illumina reads. Bottom - assembled using long Nanopore reads, showing one chromosome and five plasmids.

Chapter 2 figures:

Figure 2.1 – Schematic of the assembly correction workflows, detailing the programmes used and the type of input sequence reads for each process.

Figure 2.2 - Number of discrepant single base substitutions that remain for each correction method vs the publicly available reference genome (BA000007) using Guppy v3.2.6 (top) and Guppy v5.0.11 (bottom).

Figure 2.3 – Circos plot showing the number and position of single-base discrepancies (regardless of size) for the final draft assembly (chromosome) with Workflow 1 with Guppy v3.2.6 (A), Guppy v5.0.11 (B) and Workflow2 with Guppy v3.2.6 (C), Guppy v5.0.11 (D) when compared to BA000007. Workflow 1 – Nanopolish [Nanopore reads] (inner black ring), Pilon [Illumina reads] (middle grey ring and Racon [Illumina reads] (outermost grey ring).

Workflow 2 – 4x rounds of Racon [Nanopore reads] (inner black ring), Medaka STEC trained model for v3/v5 respectively [Nanopore reads] (second most inner grey ring), Pilon [Illumina reads] (second most outer grey ring and Racon [Illumina reads] (outermost grey ring).

Also showing on each diagram the relative positions of the non-*stx*-encoding prophages (Grey), prophage like regions (Blue), *stx*-encoding prophages (Red) and the Locus of Enterocyte Effacement (Green).

Figure 2.4 – Precision Vs Recall of variant calling for an array of consensus ratio cut-offs and pre-masking strategies including masking positions annotated as ‘Sakai phage’ (‘SP’) and positions that are ambiguously self-mapped (‘Self’) with simulated Illumina FASTQs from the reference genome. Performed on case B.

Figure 2.5 - F1 Score for an array of consensus ratio cut-offs and pre-masking strategies including masking positions annotated as ‘Sakai phage’ (‘SP’) and positions that are ambiguously self-mapped (‘Self’) with simulated Illumina FASTQs from the reference genome.

Figure 2.6 – Maximum likelihood tree, of a “soft core” alignment of 4,475 genomes showing the tree lineages (I [pink], I/II [yellow] and II [blue]) of STEC (Clonal Complex 11). Also showing where Nanopore and Illumina sequencing data is placed within the tree for each of the two cases. All methylated positions and prophage regions have been masked. Values represent the SNP differences between the Illumina and Nanopore data for both cases.

Chapter 3 figures:

Figure 3.1 – Maximum-likelihood phylogenetic tree of 105 CC11 STEC O157:H7 genomes including n=17 (labelled) publicly available genomes and domestic UK nanopore sequenced genomes produced during this project. Sub-lineages are coloured as follows: I/II, grey; Ia, yellow; Ib, orange; Ic, red; IIa, purple; IIb, blue; IIc, green. Scale bar indicates substitutions per site of a soft-core alignment of variants from each sample in the phylogeny.

Figure 3.2 – Easyfig pairwise alignment representing the chromosome and prophage content within the samples sequenced in this study (in descending order PT4, PT2, PT49, PT21/28, PT8 and PT34). *stx2a*-encoding, *stx2c*-encoding and *stx1*-encoding prophages are highlighted in red, orange and yellow, respectively. Prophage-like elements are coloured blue, and the locus of enterocyte effacement is shown in green. Finally, the non-*stx*-encoding prophages are coloured black. Scale equates to 1 million base pairs.

Figure 3.3 – Mid-rooted tree of *stx*-encoding prophages based on Jaccard distance produced from Mash. Strains are annotated with strain ID, length (bp), *stx* subtype and SBI. Strains are coloured by sub-lineage: green, Ia; red, Ic; blue, I/IIa; grey, I/IIb; orange, IIa; black, IIb; purple, IIc. Scale bar indicates Jaccard distance. * denotes the non-domestic strain sequenced in this project 194195.

Figure 3.4 – Easyfig alignment comparing the *stx1a*-encoding prophages from 644 (x2), 180 (x2), 350, E30288, EDL933 and Sakai. Arrows indicate gene directions. *stx* genes are shown in red; recombination/replication genes are shown in light blue; regulation-associated genes are shown in dark blue; effector genes are shown in pink; phage structure- and lysis-associated genes are shown in light and dark green, respectively; tRNAs are shown as purple lines; finally, hypothetical genes are shown in grey.

Figure 3.5 – Easyfig alignment comparing the *stx2c*-encoding prophages from all samples in the study, including two *stx2a*-encoding prophages that are associated with a *stx2c*-encoding prophage structure in red (315126 and E45000). Arrows indicate gene directions. *stx* genes are shown in red; recombination/replication genes are shown in light blue; regulation-associated genes are shown in dark blue; effector genes are shown in pink; structure- and lysis-associated genes are shown in light and dark green, respectively; tRNAs are shown as purple lines; finally, hypothetical genes are shown in grey.

Figure 3.6 – Two Easyfig alignments comparing the *stx2a*-encoding prophages from E45000 with E116508 (above) and 155 and 267849 (below), in descending order. Arrows indicate gene directions. *stx* genes are shown in red; recombination/replication genes shown in light blue; regulation-associated genes are shown in dark blue; effector genes are shown in pink; structure- and lysis-associated genes are shown in light and dark green, respectively; tRNAs are shown as purple lines; finally, hypothetical genes are shown in grey.

Chapter 4 figures:

Figure 4.1 – Maximum-likelihood phylogenetic tree of the four outbreak samples within this outbreak and the seven most closely related isolates in the UKHSA archive to show context. The four outbreak cases (A–D) sequenced by both Illumina and Nanopore technologies are shown.

Figure 4.2 – The chromosomal loci of the 15 prophages (16 in sample D). Red prophages indicate *stx* gene encoding prophages. Blue indicates prophage like elements. Green indicates the locus of enterocyte effacement (LEE).

Figure 4.3 – Unrooted neighbour-joining tree based on Jaccard distances showing prophages from Cases A - D with prophages from BA000007 (Sakai). Grouped by green; prophages shared in Cases A - D and Sakai, yellow; Sakai only and red; prophages unique to Cases A - D.

Figure 4.4 – Easyfig alignment of prophage 2 in all cases order descending A–D, detailing a small inversion from samples A and B relative to samples C and D (light green). Arrow direction also indicates gene direction.

Figure 4.5 – Easyfig alignment of prophage 8 in all cases order descending A–D. Showing a substitution in sample D. Arrow direction also indicates gene direction.

Figure 4.6 – Easyfig alignment showing the chromosome and loci of prophages in all samples sequenced in this study. *Stx*-encoding prophage, Red; Prophage-like region, Blue; Locus of Enterocyte Effacement (LEE), Green and other non-*stx*-encoding prophages, Black.

Figure 4.7 – Unrooted neighbour-joining trees of Jaccard distances showing prophages from samples sequenced in this study with prophages from BA000007 (strain Sakai). Prophages grouped by green are prophages shared in samples and reference genome; red are reference genome only and yellow are prophages unique samples sequenced in this study.

Figure 4.8 – Neighbour joining tree based on Jaccard distances of *stx*-encoding prophages of publicly available samples and the outbreak samples sequenced in this study. Prophages are coloured by sub-lineage of STEC O157:H7. Sub-lineage Ia, Green; Ib, Yellow; Ic, Red; I/Ia, Blue; I/IIb, Grey; IIa, Orange; IIb, Black and IIc, Purple.

Figure 4.9 – A maximum-likelihood phylogeny showing both Illumina derived and Nanopore derived SNP-typing results for samples sequenced in this outbreak. Also showing year of isolation and sample source.

Figure 4.10 – Easyfig alignment showing the chromosome and loci of prophages in all samples in the outbreak in question. *Stx*-encoding prophage, Red; Prophage-like region, Blue; Locus of Enterocyte Effacement (LEE), Green and other non-*stx*-encoding prophages, Black.

Figure 4.11 – Easyfig alignment of the four structural variation events within the prophage content in the outbreak samples. Sample 591122 is acting as the reference in each case. Arrows indicate gene directions. *stx* genes are shown in red; recombination/replication genes are shown in light blue; regulation-associated genes are shown in dark blue; effector genes are shown in pink; structure- and lysis-associated genes are shown in light and dark green, respectively; tRNAs are shown as purple lines; finally, hypothetical genes are shown in grey.

Figure 4.12 – Unrooted neighbour-joining trees of Jaccard distances showing prophages from samples sequenced in this study with prophages from BA000007 (strain Sakai). Prophages grouped by green are prophages shared in samples and reference genome; red are reference genome only and yellow are prophages unique samples sequenced in this study.

Figure 4.13 – Neighbour joining tree based on Jaccard distances of *stx*-encoding prophages of publicly available samples and the outbreak samples sequenced in this study (labelled with * preceding strain ID). Prophages are coloured by sub-lineage of STEC O157:H7. Sub-lineage Ia, Green; Ib, Yellow; Ic, Red; I/Ia, Blue; I/IIb, Grey; IIa, Orange; IIb, Black and IIc, Purple. Outbreak samples' annotations, a, race participant; b, household case; c, geographically distinct and isolated in 2018; d, geographically distinct and isolated in 2017.

Figure 4.14 – Easyfig alignment showing an outbreak exemplar (591229) *stx2a*-encoding prophage comparing to *stx2a*-encoding prophages which share the *sbcB* SBI from STEC CC11 sub-lineages IIa (E45000) and IIb (315176). Arrows indicate gene directions. *stx* genes are shown in red; recombination/replication genes are shown in light blue; regulation-associated genes are shown in dark blue; effector genes are shown in pink; structure- and lysis-associated genes are shown in light and dark green, respectively; tRNAs are shown as purple lines; finally, hypothetical genes are shown in grey.

Figure 4.15 – Maximum-likelihood phylogeny showing both Illumina derived and Nanopore derived SNP-typing results for samples sequenced in this outbreak.

Figure 4.16 – Easyfig alignment showing the chromosome and loci of prophages in all samples in the outbreak in question. *Stx*-encoding prophage, Red; Prophage-like region, Blue; Locus of Enterocyte Effacement (LEE), Green and other non-*stx*-encoding prophages, Black.

Figure 4.17 – Unrooted neighbour-joining trees of Jaccard distances showing prophages from samples sequenced in this study. Prophages grouped by green are prophages shared in all samples; red are unique (only found in a single sample) and yellow are prophages which are shared by some but not all of samples.

Figure 4.18 – Unrooted neighbour-joining trees of Jaccard distances showing prophages from samples sequenced in this study with prophages from BA000007 (strain Sakai). Prophages grouped

by green are prophages shared in samples and reference genome; red are reference genome only and yellow are prophages unique samples sequenced in this study.

Figure 4.19 – Neighbour joining tree based on Jaccard distances of *stx*-encoding prophages of publicly available samples and the outbreak samples sequenced in this study (labelled with * preceding strain ID). Prophages are coloured by sub-lineage of STEC O157:H7. Sub-lineage Ia, Green; Ib, Yellow; Ic, Red; I/IIa, Blue; I/IIb, Grey; IIa, Orange; IIb, Black and IIc, Purple.

Figure 4.20 – Maximum-likelihood phylogeny showing both Illumina derived and Nanopore derived SNP-typing results for samples sequenced in this outbreak.

Chapter 5 figures:

Figure 5.1 – Circos plot showing the 0.65Mbp chromosomal inversion between samples 432301 and 438729. Also detailing prophage regions in black, prophage-like regions in blue and *stx*-encoding prophages in red.

Figure 5.2 – Circos plot showing both 1.46 Mbp (blue) and 0.45 Mbp (red) chromosomal inversions between samples 401794 and 58888. Also detailing prophage regions in black, prophage-like regions in blue and *stx*-encoding prophages in red.

Figure 5.3 – Circos plot showing both 0.309Mbp inversion and 0.248Mbp inverted duplication between samples 60202 and 60198. Also detailing prophage regions in black, prophage-like regions in blue and *stx*-encoding prophages in red.

Figure 5.4 – Circos plot showing the chromosome of sample 432301, highlighting regions of the genome (links) that are >90% sequence similarity and <5kbp in size (blue) or >5kbp in size (red). Also detailing prophage regions (tiles) in black, prophage-like regions in blue and *stx*-encoding prophages in red.

Figure 5.5 – Circos plot showing the chromosome of sample 401794, highlighting regions of the genome (links) that are >90% sequence similarity and <5kbp in size (blue) or >5kbp in size (red). Also detailing prophage regions (tiles) in black, prophage-like regions in blue and *stx*-encoding prophages in red.

Figure 5.6 – Circos plot showing the chromosome of sample 60202, highlighting regions of the genome (links) that are >90% sequence similarity and <5kbp in size (blue) or >5kbp in size (red). Also detailing prophage regions (tiles) in black, prophage-like regions in blue and *stx*-encoding prophages in red.

Figure 5.7 – Pairwise alignment of exemplar compound prophages detected in outbreak samples. Labels detail sample ID, phage number within that sample, prophage size (bp) and integration site.

Chapter 6 figures:

Figure 6.1 – STEC O157:H7 lineage I/II maximum likelihood phylogenetic tree showing the distribution of AMR determinants.

Figure 6.2 – Pairwise alignment of the sub-lineage I/II sample 446541 in which a 9kbp Tn10 intergron integrated into a prophage (located at *clpA*) on the chromosome (at positions 2,080,998 - 2,089,926) containing the AMR determinants *strA*, *strB*, *tetA* and *sul-2* (AMR genes shown in red).

Figure 6.3 – Pairwise alignment of sub-lineage Ib showing that all the AMR determinants (*strA-strB/tetA/sul-2* or *bla_{TEM-1}/strA-strB/dfrA-1/tetA/sul-1/sul-2*) were co-located on the chromosome upon a 39,466bp fragment of DNA inserted within the start of Locus of Enterocyte Effacement (AMR genes shown in red).

Figure 6.4 – STEC O157:H7 sub-lineage Ib maximum likelihood phylogenetic tree showing the distribution of AMR determinants.

Figure 6.5 – STEC O157:H7 sub-lineage IIc maximum likelihood phylogenetic tree showing the distribution of AMR determinants.

Figure 6.6 – Pairwise of sub-lineage IIc co-located AMR regions. AMR genes shown in red, mobilisation genes shown in green, other resistance positions such as heavy metal resistance shown in blue and plasmid replicons shown in yellow.

List of tables:

Chapter 1 tables:

Table 1.1 – Table detailing the common animal sources of STEC O157:H7.

Chapter 2 tables:

Table 2.1 – Table showing both the default and optimised discrepancies from optimal workflow 1 using each Nanopore read set. Showing the number of single base substitutions, indels, two base indels and indels greater than two base pairs.

Table 2.2 – Table showing both the default and optimised discrepancies from optimal workflow 2 using each Nanopore read set. Showing the number of single base substitutions, indels, two base indels and indels greater than two base pairs.

Table 2.3 – Table showing predicted CDS, rRNA and tRNA counts from the genome annotation tools Prokka, PGAP and Bakta.

Table 2.4 – Table showing the breakdown of the total number of variants of each technology against the reference genome, followed by the numbers of masked variants within prophage or methylated positions.

Table 2.5 – Table showing the two most common discrepancies in the Nanopore optimised GATK VCFs and a breakdown of the relative proportions of these transitions compared to the total number of discrepant SNPs for both cases.

Table 2.6 – Table showing the final discrepant SNPs between the Illumina (Illum) data and Nanopore (ONT) data for case B. Also shown is the base as it is in the reference, the Illumina called base and read depth at that position and the same for the Nanopore data. Finally, also included is the locus tag relative to the reference genome and the gene annotation.

Chapter 3 tables:

Table 3.1 – Summary of the UKHSA archived and publicly available strains used within this study, with their strain ID, lineage, PT, *stx*-profile, chromosome size (bp), number of prophages in the chromosome, *stx*-encoding prophages with SBI and size (bp).

Chapter 4 tables:

Table 4.1 – Table showing the pairwise distance matrix (SNPs) for both Illumina and Nanopore technologies.

Table 4.2 – Table showing the position of the variants between the two sequencing technologies for all four samples.

Table 4.3 – Table showing the size, position and integration sites of all 16 prophages (top) and all six prophage-like elements (bottom) within each of the outbreak sample chromosomes. a refers to prophages that appear to be compound prophages (i.e. two or more prophages that are sequential, with intact integrase genes). b refers to the direction of the prophage and where the Integrase gene (*intA*) is located.

Table 4.4 – Table detailing the ID of each strain, finalised assembly metrics, plasmid replicon typing, prophage counts and assembly-based accessions.

Table 4.5 – Summary of the sequenced genomes used within this study. Showing their Strain ID, CC11 sub-lineage, Phage Type, *stx* profile, the *stx*-encoding prophage associated with each sample including prophage size and SBI.

Table 4.6 – The discrepant variant calls between Illumina and Nanopore datasets for all outbreak samples.

Table 4.7 – Summary of the genomes sequenced within this study. Showing strain ID, number of contigs, chromosome size, number of prophages and the size and incompatibility group of plasmids.

Table 4.8 – The size, position and integration sites of all 16 prophages within all samples in this study, “a” refers to prophages that appear to be compound prophages (i.e. two or more prophages that are in tandem or one integrated into other, with intact integrase genes). “b” refers to the end in which the Integrase gene (*intA*) is located. “*” denotes prophages on the border of the large chromosomal rearrangement.

Table 4.9 – The size, position and integration sites of any prophage-like elements within all outbreak samples in this study.

Table 4.10 – The discrepant variant calls between Illumina and Nanopore datasets for all 12 outbreak samples.

Table 4.11 – Summary of the genomes sequenced within this outbreak. Showing strain ID, number of contigs, chromosome size, number of prophages and the size and incompatibility group of plasmids.

Table 4.12 – The discrepant variant calls between Illumina and Nanopore datasets for all outbreak samples.

Chapter 5 tables:

Table 5.1 – Table detailing compound prophages within outbreaks sequenced in this thesis. Includes number per strain, average size across respective outbreak samples, integration sites and an approximation of how many unique phage modules per compound phage.

Chapter 6 tables:

Table 6.1 – Number of isolates predicted to confer resistance subdivided by sub-lineage and the number of cases reporting recent travel (within 7 days of onset of symptoms) outside the UK.

Acknowledgements:

I would like to express my most sincere thanks to my PhD supervisors Prof David Gally, Dr Claire Jenkins and Dr Tim Dallman for their support, experience and guidance throughout this PhD process and most importantly for giving me the opportunity to undertake this journey.

I would also like to thank workplace supervisor and friend Anaïs Painset for her support, guidance and understanding.

This extends to team GIDIS (Adam, Ella, David, Vineet, Satheesh and Matt) and everyone at GBRU, UKHSA who have supported me when my PhD projects and workplace duties conflict.

I would like to thank my Mum, Alberta, Dad, Robert and Sister, Sarah without who's support this journey would not have begun.

A special thank you to my little ones, Lyra and Gabriel who provided me so many laughs when I needed them most.

Finally, my biggest thanks to my Wife, Maya for all your love, support, and most of all your patience with me.

Extra thanks to Red Bull and Monster for an almost limitless supply of caffeine needed to create this thesis!

Layman's summary:

The Shiga-toxin producing *Escherichia coli* (STEC) bacterium causes severe infections in humans resulting in life-threatening complications. The majority of STEC O157:H7 cases are sporadic and not linked to another case, however, a fifth of cases are due to outbreaks, where humans have been infected from a common contaminated food or water source or via person to person. Within the population of STEC O157:H7, we have observed that some strains are associated with more severe disease than others. It is important to understand how this severity arises within sub-lineages that historically were not associated with severe disease. In the context of public health, understanding the nature of the emergence of *E. coli* that can cause hospitalisation and death is vital for informing public health actions and responses.

This thesis illustrates the use of novel long-read whole genome sequencing technologies and bioinformatics to understand how STEC O157:H7 sub-lineages associated with severe disease emerge within the UK population and the genetic variation of STEC O157:H7 within human outbreaks where a more severe strain has emerged. The main results of this thesis show that both long-read and short-read technologies can be integrated successfully, with the appropriate optimisation, for the analysis of the genome of STEC O157:H7. Secondly, the STEC O157:H7 genome is very plastic and can be reshaped within outbreak settings. Thirdly, that the acquisition of a particular group of bacteriophages (viruses that infect bacteria) into the STEC O157:H7 genome are associated with the increase in severity of disease observed in UK domestic lineages. Fourthly, bacteriophages that reside within the STEC O157:H7 chromosome are hotspots for larger genome

rearrangements due to their similarity and can help drive the genetic diversity we observe today. Finally, a description and characterisation of the current genomic antimicrobial resistance in the UK STEC O157:H7 population.

Thesis abstract:

Background:

Shiga toxin producing *Escherichia coli* (STEC) O157:H7 causes severe gastrointestinal disease and haemolytic uremic syndrome. As a result, it has a major impact on public health both within the UK and worldwide. Regular outbreaks can occur due to consumption of a contaminated food product, water, direct animal contact or via person-to-person transmission. Whole genome sequencing and bioinformatics has revolutionised microbial typing over the past twenty years. However, what is less understood is the true genome diversity within an outbreak of STEC O157:H7.

This thesis aims to attempt to quantify and characterise the genomic diversity within UK-based outbreaks of STEC O157:H7 using the long-read sequencing platform Oxford Nanopore Technologies with the overall aim to better understand the evolution of this zoonotic pathogen.

Results:

With optimisation to variant calling thresholds, reference-based variant calling methods can be utilised for both Illumina and Nanopore sequencing data and are comparable for single nucleotide polymorphism typing of STEC O157:H7. When designing the assembly-based methodology it was demonstrated that assembly

correction was still required when using Illumina data. Nevertheless, this led to a standardised bioinformatics method for processing Nanopore-based sequences of STEC O157:H7 genomes throughout my thesis. Shiga-toxin encoding prophages harbouring the *stx2a* gene were more diverse in terms of both gene content and integration site when compared to *stx2c*-encoding prophages. Characterisation of several STEC O157:H7 outbreaks revealed that the intra-outbreak genetic diversity of the accessory genomes is more than previously expected. Multiple copies of the same *stx*-encoding prophage were confirmed in two of the outbreaks. Large-chromosomal re-arrangements were also detected within outbreak strains. The large chromosomal rearrangements all appeared to be between homologous prophages within the STEC O157:H7 chromosome and were more likely to be associated with compound prophages at the terminus of the STEC O157:H7 genome than other prophages. However, the driving mechanism behind this process either random or selective still needs to be determined *in vivo*. Antimicrobial resistance determinants were unevenly distributed throughout the population of STEC O157:H7 with sub-lineages Ib, I/II and IIc harbouring on average more AMR determinants compared to other sub-lineages. For genomes with complete assemblies and the antimicrobial resistance genes had integrated into the chromosome, those integration sites were also co-located, typically on prophage or prophage-like regions and plasmids.

Conclusions:

This thesis has highlighted the genome complexity and diversity of STEC O157:H7 that can exist at a small-scale (intra-outbreak) and large population scale that has previously been difficult to detect by short-read sequencing technologies.

In UK domestic sub-lineages associated with severe disease, the acquisition of the *stx2a*-encoding prophages appears to be independent and likely from different sources whereas *stx2c*-encoding prophages are maintained vertically.

The ability to characterise the accessory genome of STEC O157:H7 in this way is the first step to understanding the significance of these microevolutionary events.

With regards to antimicrobial resistance, domestic STEC O157:H7 strains are typically susceptible with sub-lineages associated with travel harbouring more antimicrobial resistance genes.

This thesis has demonstrated that characterisation and localisation of individual STEC O157:H7 genetic components, be that SNPs, genes, prophages, plasmids and up to chromosomes rearrangements is now possible through the use of long-read sequencing technologies and can be achieved on a high-throughput scale for application within public health microbiology.

Chapter 1:

Introduction.

1.1 Pathogenic *Escherichia coli*.

Escherichia coli was first described by Theodor Escherich in 1885 as *Bacterium coli commune* ^[1]. *E. coli* typically colonises the gastrointestinal tract of mammals, including humans ^[2]. *E. coli* is typically a commensal bacterium, however, there is a variety of *E. coli* that can cause disease. Pathogenic *E. coli* are classified into “pathotypes” grouped by the type of infection caused and the associated virulence factors.

Primarily there are two major groups, extraintestinal pathogenic *E. coli* (ExPEC) and diarrheagenic *E. coli* (DEC) or intestinal pathogenic *E. coli* (IPEC) ^[3]. The ExPEC pathotype are characterised by the ability to cause an infection outside of the human gastrointestinal tract ^[3]. The DEC (or IPEC) pathotype are defined as any *E. coli* causing intestinal infection leading to diarrhoea ^[4], and includes enteroaggregative *E. coli* (EAEC), enterotoxigenic *E. coli* (ETEC), enteroinvasive *E. coli* (EIEC), diffusely-adherent *E. coli* (DAEC), enteropathogenic *E. coli* (EPEC) and Shiga toxin-producing *E. coli* (STEC) also known as Enterohaemorrhagic *E. coli* (EHEC) or Verocytotoxigenic *E. coli* (VTEC) (named as STEC hereafter).

The EAEC are characterised by the ability for cells to aggregate and to adhere to the intestinal enterocytes in a “stacked brick” formation. The aggregation occurs using aggregative adherence fimbriae (AAFs) ^[5]. Molecularly EAEC are defined by the presence of the *aggR* gene which encodes for a transcriptional regulator which is associated with induction of the above virulence pathway and thus diarrhoeal disease ^[6].

The ETEC are characterised by the ability to produce heat-labile and/or heat-stable toxins, which are named based on if the toxin can be inactivated by heat ^[7]. The heat labile toxin is similar (80%) in structure to the cholera toxin produced by *Vibrio cholerae* ^[8]. As a result, this pathotype is known to cause a watery diarrhoea and is more commonly known as traveller's diarrhoea ^[7].

The EIEC are characterised by their ability to invade human enterocytes, this pathotype is more closely genetically related to the *Shigella* genus than all other *E. coli* and there are proposals to consider the *Shigella* genus within the *Escherichia* genus ^[9]. The EIEC harbour a type III secretion system, encoded by the plasmid based *lpaA* to *lpaH* genes, which grants this invasive ability ^[10]. EIEC can cause inflammatory colitis but otherwise causes the production of watery diarrhoea like most other DEC pathotypes ^[4].

The DAEC are defined by the appearance of colonisation during infection, that there is a diffuse pattern of adherence, hence the designation ^[11]. Also common during DAEC infection is that enterocytes are forced to produce long cell extensions that surround the DAEC cell ^[12].

The EPEC are characterised by the ability to produce “attachment and effacement” lesions at the intestinal epithelial layer of the host gut mucosa. This mechanism is encoded in a pathogenic island known as the locus of enterocyte effacement (LEE) which contains the *eae* gene encoding a protein called intimin. It is the protein intimin that mediates the attachment of the cell to the enterocyte on what is known as a “pedestal” formation ^[13].

Finally, the STEC are characterised by harbouring the prophage-encoded Shiga toxins and the ability to cause severe bloody diarrhoea which lead to further complications such as haemolytic uremic syndrome (HUS) ^[14]. Like EPEC above, STEC can also carry the LEE and as a result can also form attaching/effacing lesions.

1.2 Shiga toxin-producing *Escherichia coli* (STEC).

There are over 200 serotypes of *E. coli* ^[15] however, the most common STEC serotype is O157:H7 and as a result it has received the focus of research and surveillance strategies since its discovery in 1981 ^[16]. There are other serotypes associated with the STEC pathotype, the most common are termed “the big six” non-O157 and they are O26, O45, O103, O111, O121 and O145 ^[17].

The eponymous virulence factor of STEC is the Shiga toxin. This toxin is encoded by *stx* genes located on lambdoid-like prophages ^[18], with two *stx* genes (*stxA* and *stxB*) encoding the alpha and beta subunit of the toxin respectively ^[19]. There are two immunologically distinct subtypes of *stx*, termed *stx1* and *stx2*, with sequence similarity only about 55-60% between the two subtypes ^[19]. The two subtypes of *stx* are further characterised into three *stx1* types (*stx1a*, *stx1c* and *stx1d*) and seven *stx2* types (*stx2a* to *stx2g*) ^[19].

New *stx2* sub-types have been identified in recent years including *stx2h* ^[20] isolated from wild marmosets, *stx2i* ^[21] isolated from shrimp, *stx2j* ^[22] isolated from a human. *stx2k* ^[23] was isolated from a pig, goat and four human cases. *stx2l* ^[24] from an

unknown source and *stx2m* [25] isolated from three human cases. However, whether they are clinically relevant in humans is yet to be determined.

At the molecular level, the Shiga toxin directly binds to glycosphingolipid globotriaosylceramide (GB3/CD77) which is a cell-surface receptor. The act of binding induces internalisation via endocytosis, where it is transported to the endoplasmic reticulum and not lysosomes. Here the alpha subunit of the toxin is cleaved releasing the enzymatically active A1-subunit which then causes depurination of the 28S rRNA (Figure 1.1) thus inhibiting any protein synthesis by affected ribosomes [26, 27]. This eventually leads to the induction of the apoptosis pathway and death of the affected cells.

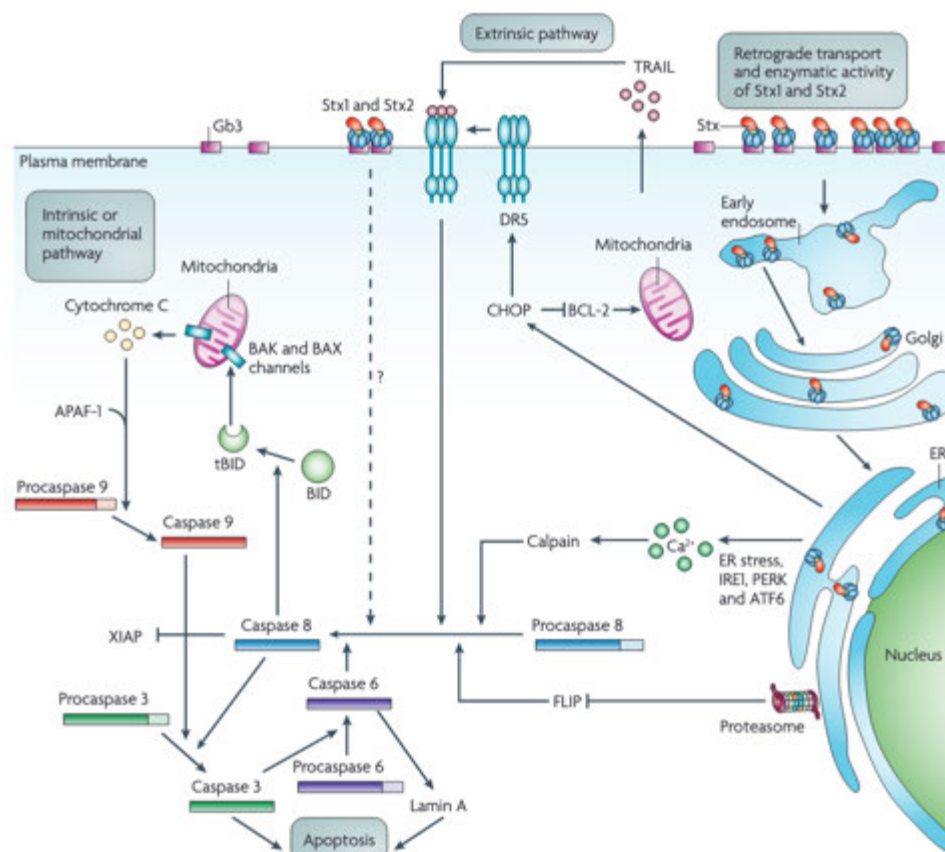


Figure 1.1 – Schematic of the binding, endocytosis, activation (cleaving) and processing of the Shiga toxin followed by apoptotic pathway induction. Adapted from Johannes and Römer 2010 [28].

Other than the Shiga toxin, STEC typically harbours an array of virulence factors to assist in adhesion, colonisation and immune evasion. STEC can also harbour the LEE as described previously, this means that STEC can also form “pedestal” like structures, known as attaching/effacing (A/E) lesions similar to EPEC [29]. The LEE is a 44.3kbp prophage-like island encoding the intimin protein (*eae*) as well as the intimin receptor (*tir*) and an array of genes (*EspA-D*, *EscA-V* and *sepZ*) that make up type III secretion systems (T3SSs) which are a set of virulence factors for injecting effectors into adjacent cells, in this case the human enterocytes [29]. Tir is injected into the host-cell membrane which then acts as the receptor for intimin [30]. The region of Tir that is situated on the intracellular surface then triggers the recruitment of actin and subsequent pedestal formation [31]. Thus, carriage and expression of *eae* and *tir* greatly increases the contact between human and STEC cells increasing the effectiveness of other T3SSs.

STEC is also known to harbour a diverse set of effector genes, which are vital for interaction with enterocytes. Some of these genes are located on the LEE (e.g. *EspF*) but many are located throughout the genome on the tail end of lambdoid prophages, known as non-LEE effectors (*Nle* genes) where Tobe, 2006 [32] have experimentally confirmed 39 different effectors. Another aspect that promotes adhesion of the STEC O157:H7 bacterium to the host cell are the flagella, where Wolfson *et al* 2020 [33] demonstrated that flagella can both create favourable surface

topologies for binding of other factors and that the flagella can interact with cytoskeletal components.

STEC can cause symptoms ranging from mild such as vomiting, abdominal cramps and abdominal pain though to severe bloody diarrhoea including haemorrhagic colitis. The infection can also progress into severe life-threatening sequela/complications including, thrombocytopenic purpura (TTP) and haemolytic uremic syndrome (HUS) in about 10-15% of cases ^[14]. HUS is characterised by rapid kidney failure, cardiac and neurological complications and can be fatal ^[34].

1.3 Shiga toxin-producing *Escherichia coli* (STEC) O157:H7.

STEC O157:H7 was first recorded as a pathogenic agent in 1982 during a multi-state foodborne outbreak associated with a fast-food restaurant chain in the USA ^[16]. In the UK, the first recorded outbreak of STEC O157:H7 was in the West Midlands in 1982-3 ^[35]. Since then, it has been detected on almost every continent and responsible for many large outbreaks of gastrointestinal disease globally ^[36,37,38].

STEC O157:H7 has a low infectious dose relative to other gastrointestinal pathogens however it is similar to the infectious dose of *Shigella* species, that being less than 700 cells ^[39] and could be as low as 10 cells ^[40]. As a result, STEC O157:H7 is easily transmissible via the faecal-oral route.

It has been determined that STEC O157:H7 is a zoonotic foodborne pathogen ^[41] and is typically transmitted via the faecal-oral route on food and water contaminated with the faeces of ruminants, who are asymptomatic shedders of the bacterium ^[42].

The bacterium colonises the terminal end of the ruminant's gastrointestinal tract at the recto-anal junction, with less evidence that STEC O157:H7 colonises elsewhere in the ruminant gastrointestinal tract ^[41].

There are many animals in which STEC O157:H7 has been isolated from including wildlife, domesticated animals, non-ruminant farm animals but the most common are ruminant animals such as cattle or sheep (Table 1) ^[42].

Group	Animal	Citation
Ruminants	Cattle	Chapman <i>et al</i> 1997 ^[43] . Milnes <i>et al</i> 2008 ^[44] .
	Sheep	Chapman <i>et al</i> 1997 ^[43] . Milnes <i>et al</i> 2008 ^[44] .
	Goats	Orden <i>et al</i> 2008 ^[45] .
Non-ruminants	Pigs	Chapman <i>et al</i> 1997 ^[43] . Milnes <i>et al</i> 2008 ^[44] .
Domesticated animals/pets	Dogs	Hancock <i>et al</i> 1998 ^[46] .
Wildlife	Deer	Ahn <i>et al</i> 2009 ^[47] .
	Boar	Sanchez <i>et al</i> 2010 ^[48] .
	Birds	Renter <i>et al</i> 2003 ^[49] . Wallace <i>et al</i> 1997 ^[50] .

Table 1.1 – Table detailing the common animal sources of STEC O157:H7 and associated literature.

Approximately 0.1% ^[51] to 18% ^[52,53] of colonised cattle make up 99% of the introduction of the pathogen into the environment ^[54]. The rate at which cattle excrete or shed STEC O157:H7 is very variable with some cattle shedding 100 CFU/g, which is the lower limit of detection, up to up to 10⁷ CFU/g of STEC O157:H7 in faeces. Those that shed at the higher end of the scale are termed “super-shedders” ^[55].

Meat products from ruminants are contaminated or during slaughter ^[56] and it is common for other animal products to be contaminated such as milk ^[57] and cheeses ^[58]. Unpasteurised products are frequently found to be the source of outbreaks of STEC O157:H7 in the UK ^[59]. Similarly raw meat products sold as pet food for domesticated animals, in particular dogs also cause outbreaks of STEC O157:H7 ^[60]. Finally, another method of transmission is direct contact with contaminated animals such as contact at a petting farm ^[61].

Contaminated water can also be a vehicle for STEC O157:H7, this has been observed in both agricultural runoff containing faeces contaminating rivers, streams, beaches ^[62,63,64] and into private water supplies ^[65,66].

Another common method of transmission of STEC O157:H7 is via contaminated ready to eat foods such as contaminated salads, leaves or vegetables ^[67] with the largest recorded STEC O157:H7 outbreak in the UK being related to contaminated leeks and potatoes in 2010 to 2011 totalling 252 cases ^[68].

Person to person transmission of STEC O157:H7 is not uncommon but usually occurs in institutional settings such as prisons, hospitals, care homes [69,70] and most commonly within childcare facilities [71,72].

1.4 Genome evolution of STEC O157:H7.

With the advent of whole genome sequencing, microbiology moved into a new realm of genomics. The first complete STEC O157:H7 genome sequenced was EDL933 in 2001, isolated from mincemeat [74]. When comparing this genome to *E. coli* K-12 [74] the first *E. coli* sequenced, they noticed that EDL933 was much larger than K-12 (genome size ~4.7Mbp), with EDL933 (genome size ~5.5Mbp) having more than 1.34Mbp of DNA not found in K-12 and 0.53Mbp found in K-12 but not EDL933 [74]. Also noted was that EDL933 contained 338 more genes than K-12. These genes were thought to be bacteriophage related suggesting EDL933 harboured more bacteriophages than other *E. coli*. They also initially reported that the *stx* gene was located on a cryptic prophage and noted that this *stx*-encoding phage was variable among other sequenced O157:H7 strains [74].

Following on later that year in 2001 a second STEC O157:H7 genome was sequenced, associated with an outbreak in Japan in Sakai caused by contaminated radish sprouts [75]. Today, this genome sequence is known as the STEC O157:H7 strain Sakai and is the most commonly used reference genome for STEC O157:H7 [75]. The Sakai reference genome (accession - NC_002695) is 5.59 Mbp in length

with a chromosome of 5.49 Mbp and two plasmids, one 92 kbp virulence plasmid (pO157) and a small 3.3 kbp plasmid. The average GC content is 50.5 % and the chromosome encodes 5,361 protein encoding regions ^[75]. Comparing to another reference genome *E. coli* K-12 ^[74], it was recorded that at least 4.1 Mbp was 98.31 % conserved between the two genomes ^[75]. The difference between the two genomes was that the Sakai strain contained many more mobile genetic elements, in particular prophages (n=18) two of which were *stx*-encoding and prophage-like elements (n=6), one of which was the LEE ^[75].

The initial genome evolutionary model of STEC O157:H7 was proposed by Feng *et al* 1998 ^[76]. They suggested an ancestral *E. coli* that contained the LEE (EPEC), was sorbitol fermenting, could express β -glucuronidase and was most likely O55:H7 serotype based on genetic distance to modern O157:H7. Then through a stepwise model firstly acquiring a *stx2*-encoding bacteriophage, followed by the change in the *rfb* cassette changing the serotype from O55 to O157, the ancestor to all contemporary STEC O157:H7 emerged. Further evolutionary steps included the acquisition of a *stx1* bacteriophage and subsequent loss of ability to ferment sorbitol. The final step involved the loss of the ability to express β -glucuronidase leading to the modern STEC O157:H7 harbouring *stx1* and/or, *stx2*, the inability to ferment sorbitol and express β -glucuronidase (Figure 2).

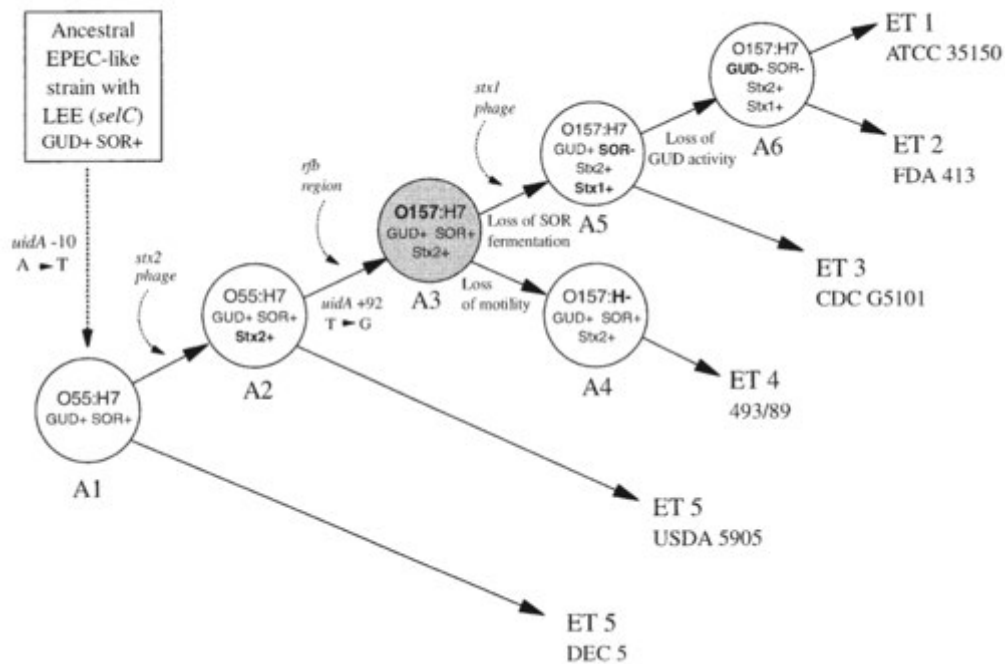


Figure 1.2 – The first stepwise evolutionary model of STEC O157:H7 proposed by Feng *et al* 1998 [76]. GUD = β -glucuronidase production and SOR = sorbitol fermenting.

The model was later updated in 2007 [77] to include putative sequence type identifiers with ST335 (associated with O55), changing to ST11 (associated with O157) at the same evolutionary step as changing serotype.

The next model of STEC O157:H7 evolution was proposed by Leopold, *et al* 2009 [78], which was the first use of whole genome sequencing to understand the acquisition single-nucleotide polymorphisms (SNPs) across the evolution of STEC O55:H7 to O157:H7. The stepwise model remains almost the same as proposed by Feng *et al* 2007 [77]. Differences included the the ability to ferment sorbitol and express β -glucuronidase were lost at the same evolutionary step and acquisition of the *stx2* bacteriophage at the start of the model, followed by loss and re-acquisition

later in the model. Leopold *et al* 2009 ^[78] also suggest that the intermediate STEC O157, *stx2* positive *stx1* negative, sorbitol fermenting and ability to β -glucuronidase express is most likely extinct. This however was in contradiction to Feng *et al* 2007 ^[77], who had found a possible example strain (LSU-61) of this intermediate step with the only aberration being that the strain did not harbour any *stx* genes.

A revised model was then proposed by Kyle *et al* 2012 ^[79], who returns to the same description as Feng *et al* 2007 ^[77], that the ability to ferment sorbitol was lost initially with loss of β -glucuronidase expression occurring later in the model. Other aspects of other model remain the same including the remaining order and detail of the steps within the model and the agreement with Leopold, *et al* 2009 ^[78], that *stx* bacteriophage was dynamic and both acquisition and loss occurred.

Kyle *et al* 2012 ^[79], were the first to use modern bioinformatics methods to generate a maximum-likelihood phylogeny of proposed genomes in each step of the model (Figure 3). This study provided more support for the revised model, based on the branching of the phylogeny and the grouping of genomes with similar features collating into clades within the tree.

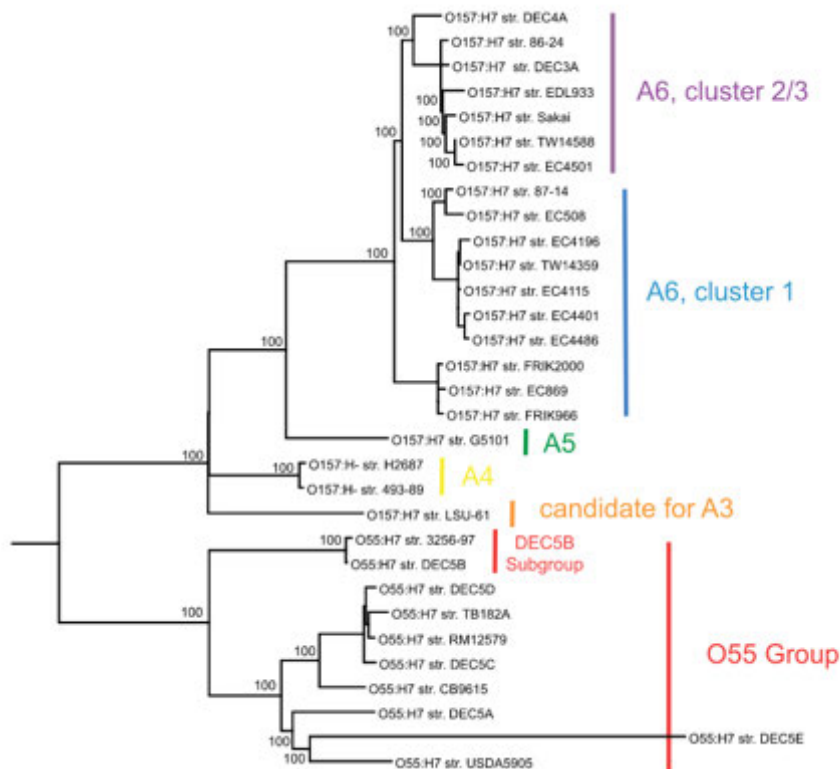


Figure 1.3 – Maximum-likelihood phylogeny grouping of STEC with similar genome features within the evolutionary model and subsequent emergence of STEC O157:H7. From Kyle *et al* 2012 [79].

The currently accepted model was proposed by Dallman *et al* 2015 [80] based on the analysis (Bayesian phylogeny) of 1,075 STEC O157:H7 genomes. It was determined that the EPEC O55 ancestor was in fact an STEC already harboured the *stx2c* gene and the initial divergence occurred following the alterations to the O-antigen cassette thus changing the serotype (O55:H7 to O157:H7), approximately 400 years ago. They also described how the population of modern day STEC O157:H7 was divided into three distinct lineages, I, I/II and II in which each lineage was derived originally from lineage II (*stx2c*) which Shaaban *et al* 2016 [81] later demonstrated. Dallman *et al* 2015 [80], This was predicted to occur 175 years ago (95% highest posterior

density, ± 14 years) a significantly shorter time-period than the previous estimates of 400 years ago ^[82] or 2500 years ^[78]. Dallman *et al* 2015 ^[80] were also the first to demonstrate the loss and acquisition of multiple Shiga-toxin subtypes within each STEC O157:H7 lineage and sub-lineage. They concluded that the *stx* bacteriophage gain and loss is dynamic but must be under a selection pressure to be retained within the STEC O157:H7 reservoir.

1.5 Detection and typing of STEC O157:H7.

Traditional culturing methods of STEC O157:H7 are based on the use of selective media such as cefixime tellurite sorbitol MacConkey (CTSMAC) agar. STEC O157:H7 having lost the ability to ferment sorbitol produce colonies with no colour (colourless) when cultured at 37 °C on this media. Most strains of *E. coli* in the human gut ferment sorbitol and appear bring pink on CTSMAC agar. STEC O157:H7 as it is cefixime resistant and so the presence of cefixime selects for STEC O157 and prevents the growth of other organisms ^[83,84].

Serological typing or serotyping is an antisera agglutination method developed to determine the somatic (O) antigen of *E. coli* ^[85]. Results are determined by the reaction (agglutination) of a fixed culture of the target strain with a known antiserum. However, due to the storage requirements and maintenance of each of over 200 different types of antisera and the occurrence of non-specific cross reactions ^[86] this method was difficult to scale and implement in laboratories and lacked robustness and reproducibility.

In 2004 a rapid real-time polymerase chain reaction (PCR) was designed for the design on the main STEC serotypes including O157 by detection of the *stx1*, *stx2*, *eae* genes and the unique O157 section of the somatic antigen encoding gene (*rfbeO157*) [87]. This assay can be used to confirm that a cultured isolate is indeed STEC O157 and to determine the primary virulence content of the isolate with a rapid turnaround, in only a few hours.

Another traditional sub-typing method of STEC O157:H7 is phage typing. This is a process where a panel of sixteen bacteriophages is spotted on to a lawn of growth of STEC O157:H7 on an agar plate. A phage type (PT) is determined based on the pattern of bacteriophage lyses [88,89]. Phage-typing has played an important role in the surveillance and epidemiological analysis of STEC O157:H7 by increasing the resolution of typing, which can be used to help inform public health actions and interventions.

Molecular subtyping methods were developed to determine a higher resolution of sample relatedness. An example of this was pulsed-field gel electrophoresis (PFGE) [90], where genomic DNA of the query sample is digested by enzymes, and the resulting fragments are separated within a gel based on size to produce a set of “bands”. Isolates which are more related exhibit similar band patterns. The more fragments that are positioned in the same place in the gel, the closer the strains are in terms of genetic relatedness. This information was then used to help inform epidemiological investigations.

In the UK since 2015, the previous typing methods have been supplemented with whole genome sequencing (WGS) which is the process of sequencing all the genomic information within a given sample [91]. WGS gathers all genetic information

on a bacterium making it possible to derive all typing results from that data *in silico* using bioinformatics. This means that the typing results derived from a variety of laboratory tests, such as PCR-based virulence profiles, serotype and strain relatedness, can all be obtained from a single assay. Whole genome sequencing is currently being used to speciate, serotype, assign a multi-locus sequence type (MLST) (typing of seven housekeeping genes) ^[92], to detect and call the *stx* subtype ^[93], to variant call and perform SNP typing ^[94] on STEC O157:H7.

1.6 Population structure of STEC O157:H7 in the UK.

The MLST sequence type (ST) most frequently assigned to STEC O157:H7 is ST11 (when using the Achtman scheme). Within ST11 the population is divided into three main distinct lineages termed as I, I/II and II, based on several phylogenetic and genome relatedness methods ^[80,91,95,96,97,98,99]. The three main lineages of I, I/II, II can be further subdivided into eight sub-lineages. Lineage I can be divided into three sub-lineages, Ia, Ib, Ic. Lineage II can be divided into three sub-lineages, IIa, IIb, IIc. I/II can be divided into two sub-lineages, I/IIa and I/IIb ^[80].

Since detection and inception of phage-typing within the UK for STEC O157:H7 for national surveillance, there has been different phage-types increasing and becoming the predominant type and subsequently decreasing and being replaced by another. During the earliest recorded results in the 1980's until the start of the 1990's, the predominant phage types were PT1, PT2 and PT49. While notification of PT1 and PT49 decreased, PT2 notifications increased ^[100]. From 1990 to 1997, PT2 was by

far the most commonly typed O157:H7 however, post 1997 until today PT2 has fallen almost annually for every year on record. The dramatic fall of PT2 resulted in a replacement event with PT21/28 rapidly became the predominant type (forming a large part of sub-lineage Ic). Concurrently, we observed PT8 rising slowly from 1998 to the early 2010s ^[100]. In the recent years PT21/28 declined and was replaced with PT8 (forming a large clade in sub-lineage IIc) as the predominant type from 2012 onwards ^[100].

1.7 Epidemiology of STEC O157:H7 in the UK.

During the years of 1983 to 2012 the crude incidence of STEC O157:H7 in England and Wales was 1.49 to 1.81 (average 1.65) cases per 100,000 persons-years equating to an average of 887 cases per year ^[100]. This is compared to all serotypes of STEC which had an incidence of 1.74 to 1.86 (average 1.81) cases per 100,000 persons-years for the years of 2009 to 2012 ^[102]. There are only two more countries with a documented higher rate of STEC O157:H7 per population than the UK those being Ireland and New Zealand (>1.80 up to 4.28 cases/100,000) ^[102]. The incidence of STEC O157:H7 is not equality distributed by age, sex and geography of the UK. Firstly, STEC O157:H7 has a higher incidence in females (55.3%) than males (43.0%) (1.7% unreported) and almost 50% of cases were from children aged 15 or less (45.1%) with a large proportion (22.8%) of cases being within the ages of one to four ^[100]. Disease severity and progression to HUS and other complications are not equally distributed across the age groups with progression to severe disease and

HUS occurring mainly in children under the age of 15 and increasing relative counts as the age group lowers to the highest of almost 12-15% of cases of birth to four-year-olds. The English regions with higher incidence of STEC O157:H7 are the North East, North West and South West of England and the lowest incidence in London [100].

Microbiologically, disease severity is also unequally distributed across both the phage types, *stx* subtypes and sub-lineages of STEC O157:H7 with *stx* subtypes showing a combinational effect to disease severity. The three most common *stx* subtypes in STEC O157:H7 in the UK are *stx1a*, *stx2a* and *stx2c* [80,103]. The subtype *stx2a* or *stx2a* and *stx2c* are more associated with severe symptoms such as bloody diarrhoea, progression to HUS and fatalities. Whereas *stx2c* presents in the majority of milder illness cases that were asymptomatic or presented with non-bloody diarrhoea [103]. However, any combination of *stx1a* and other subtype typically lead to more severe disease suggesting that both *stx2a* and *stx1a* lead to a more severe disease than *stx2c*. Keeping in mind a bias that gastrointestinal disease reporting are biased toward the more severe clinical cases [103].

The variation of severity of disease by sub-lineage of STEC O157:H7 is related to the *stx* subtype that is harboured by strains in that sub-lineage. Sub-lineages I/II and Ic both have strains containing *stx2a* or *stx2a/stx2c*. Within sub-lineage Ic strains with *stx2a* and *stx2a/stx2c* were associated with severe disease progression, however in sub-lineage I/II those with *stx2a* were more likely to be associated with severe disease progression than *stx2a/stx2c* [80,103]. Another aspect to consider is the variation in disease severity within a particular lineage over time. *Stx*-encoding bacteriophages are mobile and can be acquired and lost throughout sub-lineages

thus changing the risk associated with severe disease. An example of this is sub-lineage IIb where the ancestral strain harboured *stx2c* and the rate of bloody diarrhoea was 22.6% ^[103]. Between the year 2000 and 2010 the *stx2c* prophage was lost and as a result rate of bloody diarrhoea fell to 16.7%. More recently, over the last decade we observed the acquisition of *stx2a* and the rate of bloody diarrhoea linked to this sub-lineage sharply rose to 54.3% ^[103].

The reasoning for the variation in clinical severity with *stx* subtype is not yet known but may include the interplay of prophage content and the structure of the prophages where the *stx* genes are located, larger genetic interactions not only with the prophages but across the wider accessory genome and finally host susceptibility. For example, some individuals could be more susceptible to infection than others.

Outbreaks of STEC O157:H7 are common and typically range from 0 – 25 annually in the UK ^[100]. There have been numerous outbreaks associated with each of the known vehicles and methods of transmission ^[104]. Some of the outbreaks include drinking raw milk ^[59,105] and milk pasteurisation failures ^[57]. Contaminated meat products ^[106] including mince ^[56]. Contaminated salad products ^[107] and vegetables, which also accounts for the largest recorded outbreak in the UK, including 252 cases related to contaminated raw leeks and potatoes ^[68]. Direct contact with shedding animals, domestic animals or at petting farms ^[61,64].

Person to person transmission can occur in facilities like prisons, hospitals, care homes ^[69,70] and most commonly within childcare facilities ^[71,72].

1.8 Bacteriophages and prophages associated with STEC O157:H7.

Bacteriophages are viruses that can infect and cause cell death of bacteria. Many bacteriophages have a typical life cycle where free bacteriophages infect a given bacterium and exhibit either the lytic or lysogenic phase. The lysis phase is where the bacteriophage replicates using the bacterium host machinery and lyses the bacterium leading to cell death. The lysogenic phase is where the bacteriophage's DNA integrates with the host genome or DNA to await activation to the lytic cycle. In the lysogenic phase the bacterium replicates its genome as normal and thus also replicates the integrated bacteriophage which is now known as a prophage as shown in figure 1.4 [108].

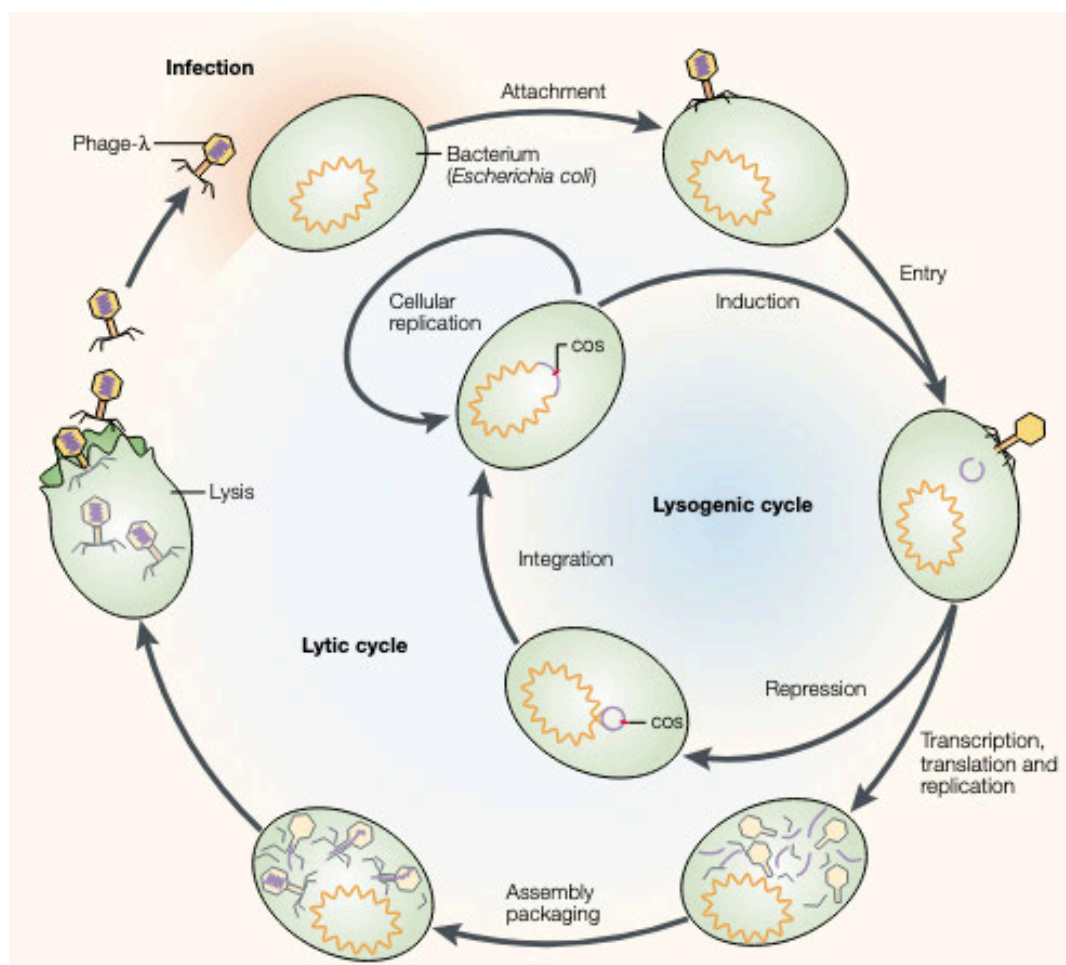


Figure 1.4 – Bacteriophage replication showing both lytic and lysogenic cycles of Phage- λ (from Campbell 2003 ^[108]).

There are advantages and disadvantages of each path on the life cycle for the bacteriophage. Lysogeny can provide benefits to both the bacteriophage and the host by bringing new genes and integrating them into the chromosome of the host bacterium and potentially providing a selective advantage to the bacterium in response to environmental pressures ^[109]. Acquisition of prophages that are on average are 45 kbp in size, is a considerable burden on the host for replication where in an energy deficient environment is deleterious for the host ^[109]. Equally, induction to the lytic pathway in situations where a specific host bacterium is under stress, provides an exit mechanism for the prophage. However, if the environment is stressful to all the potential hosts, initiating the lytic pathway could be a disadvantage.

Lysogeny also carries the risk of “Grounding” this is where a genetic event such as homologous recombination or methylation occurs to the prophage genome which removes its ability to excise forcing them to remain in the host genome permanently ^[109]. Once grounded prophages are then susceptible to the normal deleterious deletions and erroneous replications that can be afflicted on the chromosome of the host by environmental pressures ^[109]. When a genome harbours multiple prophages of a similar type, it provides more homologous sequences and thus can drive host genome homologous recombination leading to more genetic diversity of the host ^[109].

STEC O157:H7 has a large number of prophages per genome when compared to other gastrointestinal bacterial pathogens with the prophage and prophage-like elements making up approximately 13.5% of the genome ^[110] and upper estimates of 16% ^[111]. STEC O157:H7 strain Sakai was found to harbour eighteen prophages and six prophage like regions on the chromosome shown in figure 1.5 ^[75].

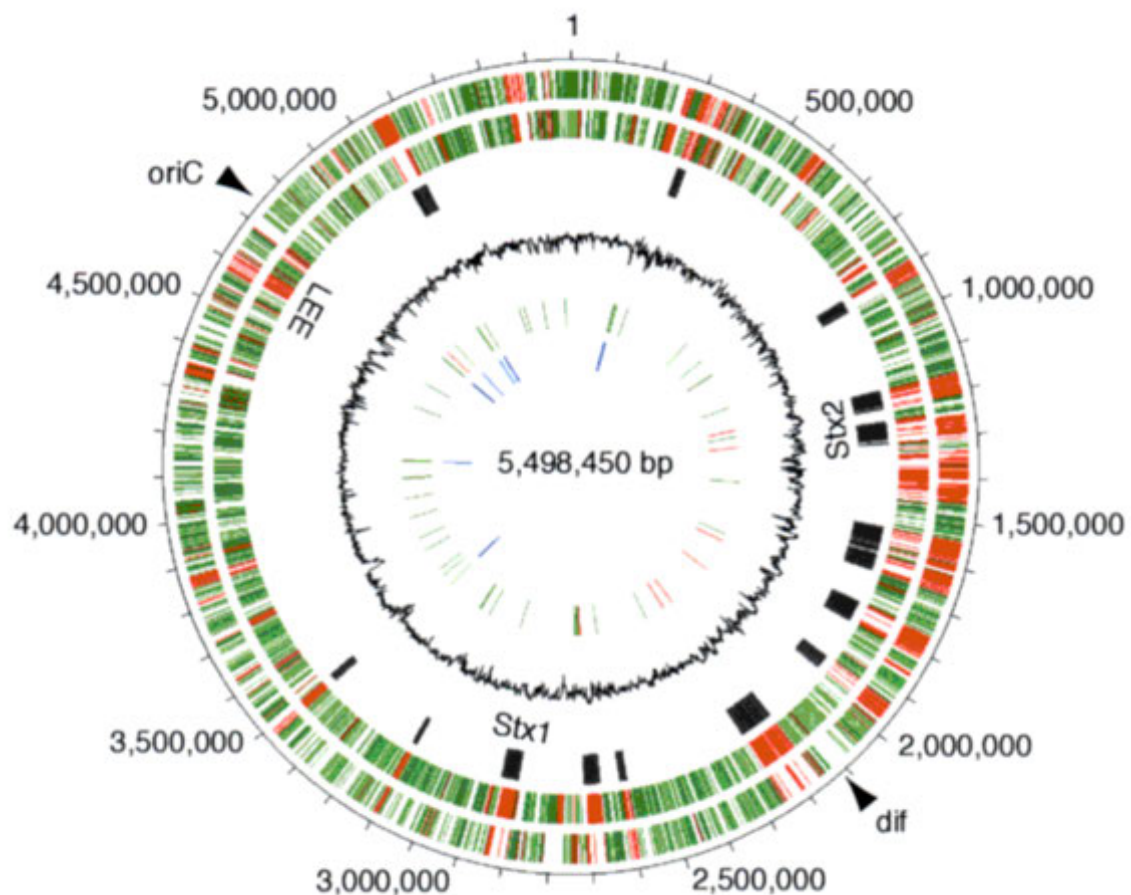


Figure 1.5 – Chromosome (~5.5 Mbp) of STEC O157:H7 strain Sakai with inner ring showing black blocks which equate to prophages and their relative positions on the chromosome (from Hayashi *et al* 2001 ^[75]).

Another key consideration is that prophages within STEC O157:H7 can be mobile, meaning that a prophage can move from one chromosomal locus to another. There have been numerous studies that have identified different loci of prophage integration ^[112,113,114,115,116,117] contradicting the evidence of the previous assumption that prophages integrate at a single locus.

However, it has been shown that not all prophages within STEC O157:H7 are active but rather some are defective/grounded and cannot enter a lytic state. Asadulghani *et al* 2009 ^[111], showed that that 17 out of 18 prophages in the Sakai reference genomes contained defective genes, however, at least 9 could still excise themselves from the chromosome and replicate. These inducible prophages were also grouped into three categories: (i) those prophages that could be induced spontaneously, (ii) those induced with an inducing agent such as MMC or UV light and (iii) those induced by regional replication of other prophages, i.e. the replication of another prophage triggers the induction of another. This is extremely important as many ends of prophages carry non-LEE effectors (*nle* genes) which are associated with progression of more severe disease ^[116].

1.9 Shiga-toxin encoding bacteriophages.

A subset of prophages found in STEC O157:H7 encode the archetypical Shiga-toxins. The *stx*-encoding prophages are lambdoidal in structure, meaning they share much of the same gene content as bacteriophage lambda which was discovered in

the 1980's ^[18,118]. Later, the structure of *stx2* and *stx1*-encoding prophages were shown to have some conserved regions of the bacteriophage genome but other regions showed variation which led to the *stx2* and *stx1*-encoding prophages genomes to be described as modular ^[113,119]. As more *stx*-encoding bacteriophage genomes were sequenced it quickly became apparent that they ranged in their size, from 47kbp to greater than 70kbp ^[120]. The site of *stx*-encoding bacteriophage integration (SBI) also varies, not only in STEC O157:H7, but across all STEC serotypes with some sites showing specificity to one site or another such as *argW* or *yecE* always containing *stx2*-encoding prophages or *yehV* for *stx1*-encoding prophages. However, some SBIs have been observed to harbour both types of *stx*-encoding prophages such as *wrbA* ^[116]. Shaaban *et al* 2016 ^[81] confirmed that the *stx2c*-encoding prophage's SBI was *sbcB* and demonstrated that this prophage was maintained at this location throughout the STEC O157:H7 population.

The expression of Stx is directly related to the bacterial SOS response. Any form of DNA damaging agent such as mitomycin C (MMC), ultraviolet light or some antibiotics will cause the activation of the *recA* gene which is one of a set of genes responsible for the repair of DNA following such damage in bacteria. However, activated RecA also has the secondary effect of inactivating prophage repressor proteins and activating prophage proteins driving the lytic cycle. This results in expression of prophage structural proteins and the sequential *stx* genes as well ^[120]. As a result, the majority of temperate phages (prophages) have evolved a trigger to move to a lytic state when the bacterium engages an SOS response into order to ensure its own survival ^[121]. For *stx*-encoding prophages, this leads to a complication for treatment of STEC O157:H7 infection as Wong *et al* 2000 ^[122], demonstrated that

there is a significantly higher risk of children infected with STEC progressing to HUS when given antibiotics than those who are not ^[122,123]. This is due to some antibiotics causing DNA damage and thus triggering the SOS response leading to Stx production and when cell lysis occurs the concentrated Stx is released.

1.10 STEC O157:H7 genome plasticity.

Multiple prophages integrated into the chromosome sharing a high sequence similarity provides hotspots for homologous recombination. Investigating this further, Iguchi *et al* 2006 ^[124], were able to induce multiple spontaneous recombination in another STEC reference genome STEC EDL933 and noted that all endpoints of the recombining DNA were linked to the prophages within the chromosome, the largest of which was 1.4 Mbp ^[124]. It has been suggested that insertion (IS) elements ^[125,126] and prophages ^[124,125] are responsible for these inversions. Either way, any homologous sequence, either an IS element or prophage, may act as anchors increasing the opportunity for recombination in STEC O157:H7. It was also noticed that these described recombination events were not limited to non-*stx*-encoding prophages. It has been demonstrated *Stx*-encoding prophages are also involved, where induction of STEC O157:H7 strain Sakai produced a recombination event crossing the two *stx*-encoding prophages Sp5 and Sp15 amounting to a 2 Mbp inversion effectively inverting half the chromosome ^[111].

Optical mapping and PCR of the edges of the inversion, have also corroborated the existence of these events ^[127]. They have not been fully explored due to either limitations in traditional sequencing technologies being too expensive, or next-generation sequencing being confounded by the homologous sequences between prophages in STEC. As a result, large recombination events have not been detected and fully characterised. It is vital that these events are explored as they clearly play an important role in not only the evolution of the STEC genome but also the genomic architecture of prophages within the chromosome of STEC. Furthermore, the repertoire of STEC virulence genes is frequently located at the terminal edge of prophages which are also beside homologues regions between prophages and likely to be involved within these recombination events.

1.11 Impact of whole genome sequencing on UK national surveillance of STEC O157:H7.

DNA sequencing technologies have rapidly developed and progressed over the past 20 years from initial Sanger sequencing through to single-molecule real-time sequencing (SMRT). By 1977 Sanger sequencing or sequencing using dideoxynucleotides was being used to determine each nucleotide in a step wise format on an electrophoretic gel and this method was used to sequence the first genomes ^[128]. In recent decades, more high-throughput methods were developed including methods such as sequencing-by-synthesis ^[129] and pyrosequencing. Sequencing-by-synthesis WGS platforms are now make up the highest proportion of whole genome sequencing platforms on the market. WGS was introduced in the

public health institution of the UK, formally Public Health England, now UK Health Security Agency. Part of the remit of UKHSA is to conduct surveillance of pathogens that have an impact on clinical and public health. Whole genome sequencing and bioinformatics is now the routine method of characterising, typing and outbreak detection of STEC O157:H7 ^[91].

The strain resolution that whole genome sequencing has provided in terms of SNP typing is unparalleled ^[104]. The ability to distinguish between individual SNPs has led to a paradigm shift on how STEC O157:H7 outbreaks have been characterised in the UK. The ability to include or exclude strains based on genetic distance, even if an epidemiological link does not yet exist, for outbreak detection and national pathogen surveillance, cannot be achieved with any other laboratory assay or examination. WGS has also led to the increase of tools for genotyping aspects of bacterial genomes such as detection of virulence genes, typing genes such as somatic and flagellar antigen gene detection and antimicrobial resistance genes.

By combining the above microbiological strain resolution with epidemiological data, national surveillance using WGS has led to a better understanding of the strains of STEC O157:H7 that are endemic in the UK ruminant population ^[91,99,104]. This process now guides outbreak investigations, for example we can identify whether strains causing infection are being imported to the UK on contaminated produce (non-domestic) or if associated with known domestic UK reservoirs.

With many public health institutions incorporating the use of WGS for STEC O157:H7 surveillance has led to a new ability to share sequencing data on a national and international level, which has allowed for the efficient comparison of strains involved in cross-border or international outbreaks of STEC O157:H7 ^[102,130].

Studies have demonstrated that combining whole genome sequencing with machine learning approaches enables us to associate STEC O157:H7 strains in the UK with different hosts. For example, linking individual core genome SNPs from a genome to a host such as human or bovine origins ^[131].

1.12 Novel sequencing technologies.

Long-read sequencing technologies such as Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT) form the most recent generation of DNA sequencing platforms. Focusing on Nanopore, the basis of the technology uses an electric field to pull single-stranded DNA through a biological pore that is housed in a biological membrane across which a current is passed. As each nucleotide moves through the pore, it disrupts the current across the membrane, and each different nucleotide has a different disruption based on its chemical structure. These disruptions are measured and recorded and from this, the nucleotide sequence can be inferred ^[132]. As the detection method is without proxy e.g. fluorescence, hydrogen ions or other molecular tags, it means the native DNA is sequenced. This includes any DNA modifications such as methylation that is typically lost with any sequencing process that requires a polymerase reaction ^[133].

Short-read sequencing technologies generate an issue for understanding the prophage and accessory genome content. STEC O157:H7 have relatively higher prophage content and homologous sequences compared to other bacterial pathogens. *De novo* assembly of the genome using short reads is difficult as the homologous sequences within each prophage or insertion element are larger than the reads being used to assemble the genome. This leads to ambiguity in the order or sequence in which “sections” or contigs are to be placed. This has the effect of breaking the assembly into many more contigs as the assembler cannot resolve the position of the reads between many homologous sequences ^[134] (Figure 1.6, top). This also applies to large-scale recombination as short reads would not be able to resolve the entire chromosome. When comparing contigs to a reference genome, it is not possible to infer directionality, thus you cannot tell if a particular contig is inverted or not.

Long-read sequencing technologies such as Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT) can produce reads that are typically longer than next-generation sequencing technologies, such as Illumina, Ion Torrent and 454 sequencing. As a result, during the genome assembly process, these longer reads span across homologous sequences into surrounding unique sequences and thus allows for the determination of directionality. By removing this ambiguity, it allows complete circular contigs to form (Figure 1.6, bottom). This has the added benefit of allowing researchers to interrogate the previously hidden accessory genome of STEC O157:H7 including plasmids, prophages and whole chromosomal comparisons.

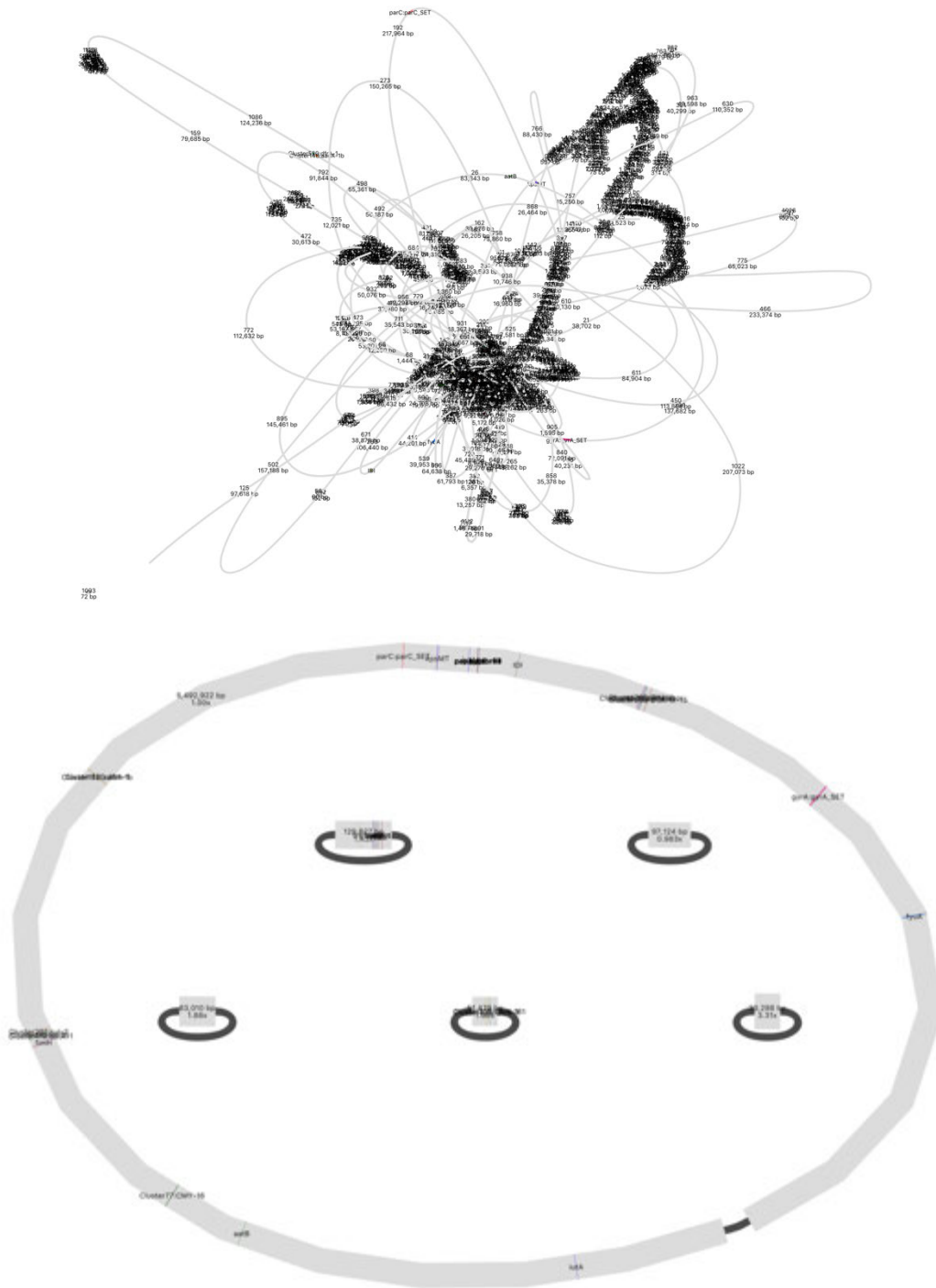


Figure 1.6 – Two *de novo* genome assemblies of the same *E. coli* isolate. Top - assembled using SPAdes with the input data being short Illumina reads, showing the non-contiguous sequences produced. Bottom - assembled using Flye with the input data being long Nanopore reads, showing one contiguous chromosome and five contiguous plasmids.

To date, only a few studies that used Nanopore or PacBio to investigate STEC have been published and given the complexity of the STEC genome, it is an excellent example candidate to showcase long read technologies. For example, one such study involved sequencing a Shiga toxin converting bacteriophage from O157:H7 [135]. Another study resolved the genomes of the major serotypes of STEC into complete assemblies [136]. While another investigated the genome of an STEC O157:H7 associated with a super-shedding steer [137]. Gonzalez-Escalona *et al* 2019 [138] showed using long read sequencing technologies to characterise STEC plasmid and virulence genes, described the integration site of a *stx*-encoding prophage and the repertoire of virulence factors on a novel plasmid in STEC serogroup O26. There however, are only two manuscripts that are directly related to STEC prophages and the use of long read sequencing, the first by Cowley *et al* 2016 [139], who used both PacBio and Nanopore sequenced isolates to investigate two closely related outbreaks, in which the core genome identified them as only a few single nucleotide polymorphisms apart, however, traditional phage typing found that they were indeed different phage types. Using long-read technologies, they were able to find a prophage (and plasmid) that was responsible for the different phage-typing results [139]. The next paper by Shaaban *et al* 2016 [81], is combined available complete STEC genomes with long-read PacBio sequenced STEC strains to investigate Shiga-toxin converting prophage diversity. They found that prophages encoding *stx2a* and *stx1a* were much more diverse than those harbouring the *stx2c* genes [81].

The use of long-read sequencing has revealed previously unknown diversity of prophages and complexity of STEC genomes. It raises questions about how this

technology can be utilised to inform surveillance of pathogens that impact on clinical and public health, and it is these aspects that I will address in this thesis as detailed in the following project aims.

1.13 PhD aims.

1. To determine if multiple sequencing technologies can be integrated for the reliable detection and analysis of STEC O157:H7 outbreaks. If so, optimise the bioinformatics processing to improve accuracy for both assembly-based and assembly-independent methodologies.
2. To detect and characterise the *stx*-encoding prophages in emerging highly pathogenic sub-lineages of the domestic UK STEC O157:H7 population.
3. To characterise the prophage and accessory genome variation of STEC O157:H7 within the context of human outbreaks.
4. To design a scalable bioinformatics method to observe and characterise large chromosomal re-arrangements in STEC O157:H7.

Chapter 2:

Integration of long and short-read
sequencing data for the analysis of STEC
O157:H7 genomes.

2.1 Chapter 2 background.

UKHSA have been using short-read Illumina sequencing for the national surveillance of STEC since 2014 and there is now a diverse database of STEC O157:H7 genomes (approximately 10,000 genomes) [91]. Implementation of a new sequencing technology, in this case long-read Nanopore sequencing, to supplement or replace current technologies, requires validation and optimisation to ensure reliability, robustness and backwards compatibility and integration with the database of genomes already sequenced. Furthermore, data processing workflows for downstream analyses for example prophage detection and comparison, need to be developed. This chapter describes the groundwork carried out for ensuring the best possible data processing workflow is in place.

Strain relatedness can be derived via several bioinformatics methodologies however, the two primary arms of this process are assembly-dependent and assembly-independent. Assembly meaning *de novo* assembly of the genome of the strain in question.

The purpose of this chapter is to discover if Nanopore and Illumina data can be made compatible using both assembly-dependent and assembly-independent methodologies to determine strain relatedness. The assembly-independent method involves alignment of Nanopore reads to a known STEC O157:H7 reference genome and calling variants based on the consensus of said reads before a pairwise comparison of genetic distance with strains in UKHSA's 10,000 STEC O157:H7 database. The optimisation of the assembly-dependent methods described in this chapter was based on attempting to re-sequence a well described STEC O157:H7

reference genome, generate a *de novo* assembly and develop models for assembly correction using the original reference as a truth set.

2.2 Chapter 2 methodology.

2.2.1 Assembly-dependent methodology.

2.2.1.1 Bacterial culture.

Strain Sakai, the well described STEC O157:H7 reference genome (public accession BA000007) ^[75], was cultured from the culture collection archive at the UK Health Security Agency (UKHSA) [known previously as Public Health England (PHE)] using 4ml of double-strength nutrient broth (Oxoid) and incubated at 37°C overnight.

2.2.2.2 DNA extraction and library preparation for Illumina platforms.

Genomic DNA was extracted from the culture using the Qiagen Qiasymphony automated extraction platform (Qiagen, Hilden, Germany) using manufactures instructions. The sequencing library was prepared using the Nextera XP kit (Illumina, San Diego, USA) for sequencing on the Illumina HiSeq 2500 (Illumina, San Diego, USA) instrument run with the fast protocol. High quality trimmed (leading and trailing trimming at <Q30 using Trimmomatic v0.27 ^[140] producing paired-end trimmed reads in a FASTQ format.

2.2.2.3 DNA extraction and library preparation for Nanopore platforms.

Genomic DNA was extracted and purified using the Fire Monkey Kit (Revolugen, Glossop UK), with minor alterations including limited vigorous mixing steps. Genomic DNA for was quantified using a Qubit and the HS (high sensitivity) dsDNA assay kit (ThermoFisher Scientific, Waltham, USA), following the manufacturer's instructions.

Library preparation was performed using the Ligation kit SQK-LSK109 (Oxford Nanopore Technologies, Oxford, UK) with >40ng/ μ l input DNA. The prepared library was loaded onto a FLO-MIN106 R9.4.1D flow cell (Oxford Nanopore Technologies) and sequenced using the MinION (Oxford Nanopore Technologies, Oxford, UK) for 48 hours all using manufactures instructions.

2.2.2.4 Bioinformatics processing including *de novo* genome assembly.

Sequencing data was produced in a raw FAST5 format and was subsequently basecalled using both Guppy v3.2.6 and Guppy v5.0.11 using the high accuracy (HAC) model (Oxford Nanopore Technologies) into a single FASTQ format. Run metrics were generated using Nanoplot v1.8.1 ^[141]. The y-adapter from the sample's reads were trimmed, and chimeric reads split using Porechop v0.2.4 ^[142] under default parameters. The trimmed reads were filtered using Filtlong v0.2.0 ^[143] with the following parameters, min length=1000 bp, length_weight=10, keep percent=90 and target bases=250Mbp, =550Mbp, =825Mbp, =1100Mbp, to generate approximately 50x, 100x, 150x and 200x coverage of the STEC genome respectively with the longest scored reads. Each of the trimmed and filtered nanopore FASTQ files were assembled (*de novo*) using Flye v2.6 ^[144] using default parameters for each read set.

2.2.2.5 Assembly correction workflows.

To determine the most accurate assembly correction workflow for the draft assembly of the re-sequenced strain Sakai genome, four programs in different orders utilising either Nanopore and Illumina reads were generated. Where Nanopore reads were

used, each individual process was repeated four times for varying coverages of Nanopore reads including, 50x, 100x, 150x and 200x coverage.

The four assembly correction programs included: Nanopolish v0.11.1 ^[145] with --fix-homopolymer and --methylation-aware=dcn,dam parameters enabled based on alignments generated by Minimap2 v2.17 ^[146] and Samtools v1.9 ^[147]. Secondly, Pilon v1.22 ^[148] used with default parameters based on alignments of BWA v0.7.17 ^[149] and Samtools v1.9 ^[147]. Thirdly, Racon v1.2.1^[150] used with default parameters with alignments generated from BWA v0.7.17 ^[149] and Samtools v1.9 ^[147] for Illumina reads and Minimap2 v2.17 ^[146] and Samtools v1.9 ^[147] for Nanopore reads. Finally, Medaka v1.0.3 ^[151] used with default parameters based on alignments generated from Minimap2 v2.17 ^[146] and Samtools v1.9 ^[147].

For alignments using short reads and BWA v0.7.17 ^[149] a parallel alignment was made using the -x intractg parameter to evaluate the effect of aligning short-reads to very similar reference sequences.

Several variables were tested within the workflow including, applicability of Nanopore reads for correction only, applicability of Illumina reads for correction only, the effect of coverage of Nanopore reads on correction, the effect of multiple rounds of racon polishing both at the start and towards the end of the workflow, to determine the most time efficient order of the finalised workflow. Figure 2.1 shows a schematic of the workflows used and the selected reads for that process.

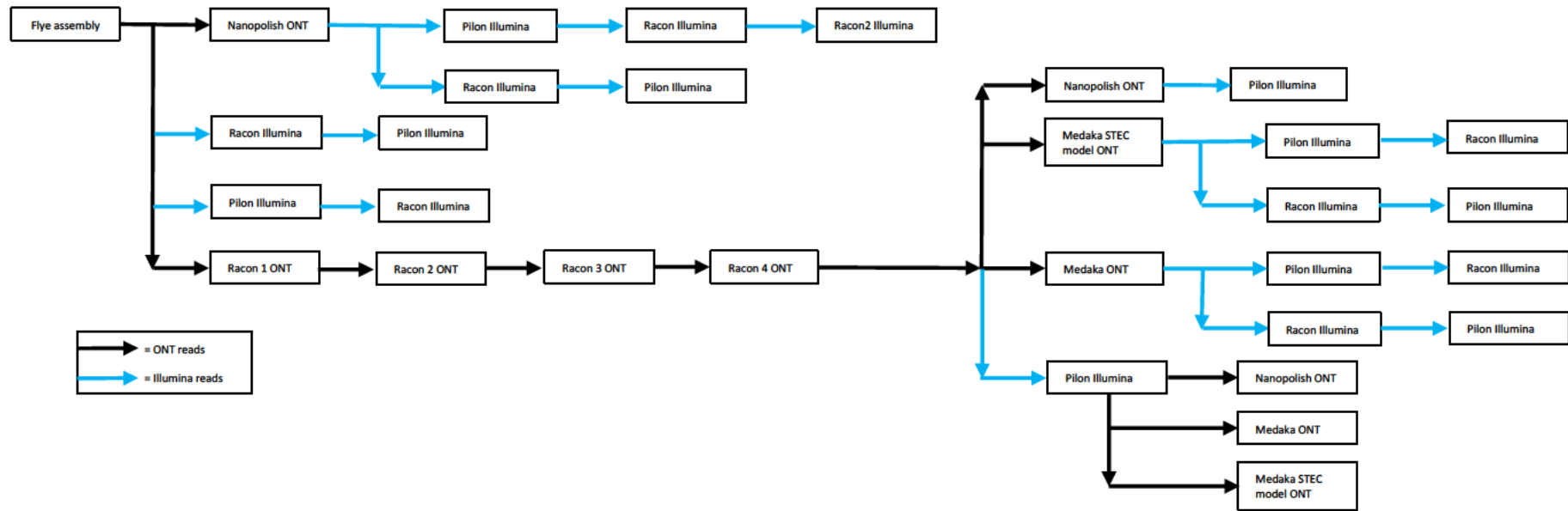


Figure 2.1 – Schematic of the tested assembly correction workflows, detailing the programmes used and the type of input sequence reads (Illumina or ONT) for each process.

2.2.2.6 Training an STEC-specific Medaka assembly correction model.

The read sets from Guppy (both v3.2.6 and v5.0.11) were each used to train a STEC O157:H7 specific Medaka model, the respective trimmed Nanopore reads were aligned to the publicly available reference for strain Sakai O157:H7 (BA000007) and aligned to the draft assembly created from Flye v2.6 ^[144] using an alignment tool from Pomoxis v0.3.6 ^[152] based on Minimap2 v2.17 ^[146]. The alignments were converted into training features using Medaka's features function followed by Medaka's training function v1.0.3 ^[153]. The STEC O157:H7 trained models were used in parallel with any Medaka steps for their respective workflows described above and in Figure 2.1.

2.2.2.7 Comparison of correction workflow outputs to truth set (Strain Sakai).

At each stage the original *de novo* draft-assembly for strain Sakai, and any polished/corrected assemblies were compared to the publicly available reference of this strain (BA000007) ^[75] using Minimap2 v2.17 ^[146] with the parameters -cx asm5 enabled. In conjunction with paf tools v2.17 ^[146], minimum alignment length set to -L 20000, to output a VCF of differences between each polished assembly and the publicly available reference genome (BA000007) ^[75].

2.2.2.8 Workflow selection and subsequent optimisation for STEC genomes.

To further the accuracy of the draft assembly produced for strain Sakai, each parameter that could affect calling of a variant and thus be used to correct the

assembly were varied and compared using Minimap2 v2.17 ^[146] as described above in section 2.2.2.7.

For Nanopolish v0.11.1 ^[145], two parameters were varied, firstly, the minimum frequency threshold (-m) of which a variant is called (increasing values of 0.1 from 0.1 through to 0.9) and secondly, the minimum depth threshold (-d) at which a variant is called (increasing values of 10 from 0 to 50) respectively. For Pilon v1.22 ^[148] using Illumina based alignments, three parameters were varied, firstly, parameters determine the minimum depth (--mindepth) to be called a variant (increasing values of 0.05 from 0.05 to 0.25). Secondly, the minimum mapping quality (--minmq) for a read to count within pileup files (increasing values of 10 from 0 to 60) and finally the minimum base quality (--minqual) to determine use within pileup files (increasing values of 10 from 0 to 40) respectively. For Racon v1.2.1 ^[150] using both Nanopore and Illumina based alignments, two parameters were varied, the maximum error rate (-e) allowed in overlaps (increasing values of 10 from 0 to 50) and quality threshold (-q) within a given window (increasing values of 0.1 from 0.1 to 0.6) respectively.

2.2.2.9 Genome annotation and visualisation.

To compare the effect different correction workflows had on the ability to call coding sequences, we processed the publicly available reference genome (BA000007), the final optimised corrected assemblies from workflow 1 and 2 and an Illumina-only assembly (assembled using SPAdes v3.2.0 ^[154]) with kmer lengths set to 31, 51 and 71), through several genome annotation tools. These tools included Prokka v1.14.6 ^[155] using --genus=Escherichia, --species=coli and --rfam enabled. The NCBI

prokaryotic genome annotation pipeline (PGAP) 2022-04-14-build6021 ^[156] using default parameters and finally Bakta v1.4.1 ^[157] using --genus=Escherichia --species=coli.

The VCF files containing the single base substitutions between any assembly from each optimised workflow produced by Minimap2 v2.17 ^[146] was visualised in Circos v0.69 ^[158].

2.2.3 Assembly-independent methodology.

2.2.3.1 Genomic DNA extraction, library preparation and Illumina sequencing.

Bacterial isolates from two cases who were admitted to the same hospital on the same night, both progressing to HUS, were taken for genomic DNA extraction. The extraction occurred using a Qiagen Qiasymphony (Qiagen, Hilden, Germany) in accordance with the manufacturer's instructions. gDNA quantification was performed using a Qubit and the BR dsDNA Assay Kit (ThermoFisher Scientific, Waltham, MA, USA) to manufacturer's instructions. The Illumina sequencing library was prepared by fragmenting and tagging the purified gDNA using the Nextera XT DNA Sample Preparation Kits (Illumina, Cambridge, UK) to manufacturer's instructions. The prepared library was loaded onto an Illumina HiSeq 2500 (Illumina, Cambridge, UK) at UKHSA [PHE at the time] and sequencing performed in rapid run mode, yielding paired-end 100 bp reads.

2.2.3.2 Illumina sequencing data bioinformatics processing.

Illumina reads (FASTQ) were processed using Trimmomatic v0.27 ^[140] to remove bases with a PHRED score of <30 from the leading and trailing ends, with reads <50 bp after quality trimming discarded. A *k*-mer approach ^[159] was used to confirm the species of the samples. Sequence type (ST) assignment was performed using MOST v1.0 described by ^[160] using the Achtman *E. coli* MLST scheme. *In silico* serotyping was performed by using GeneFinder ^[161], which uses Bowtie v2.2.5 ^[162] and Samtools v0.1.18 ^[147] to align FASTQ reads to a multifasta containing the target genes (including *wzx*, *wzy*, and *fliC*). *Stx* subtyping was performed as described by Ashton *et al* 2015 ^[93]. Illumina FASTQ reads were aligned to the Sakai STEC O157 reference genome chromosome (NC 002695.1) using BWA MEM v0.7.13 ^[149]. Variant positions were identified by GATK v2.6.5 ^[136] UnifiedGenotyper that passed the following parameters: >90% consensus, minimum read depth of 10, Mapping Quality (MQ) ≥ 30 . Any variants called at positions that were within the known prophages in Sakai were masked from further analyses. The remaining variants were imported into SnapperDB v0.2.5 ^[94].

2.2.3.3 Genomic DNA extraction, library preparation and Nanopore sequencing.

To preserve DNA integrity for the nanopore sequencing, gDNA was extracted and purified using the Promega Wizard Genomic DNA Purification Kit (Promega, Madison, WI, USA) with minor alterations including doubled incubation times, no vigorous mixing steps (performed by gentle inversion), and elution into 50 μ L of double-processed nuclease-free water (Sigma-Aldrich, St. Louis, MO, USA).

Genomic DNA was quantified using a Qubit and the HS (High Sensitivity) dsDNA Assay Kit (Thermofisher Scientific) according to the manufacturer's instructions. Library preparation was performed using the Rapid Barcoding Kit SQK-RBK001 (Oxford Nanopore Technologies, Oxford, UK) with each sample's gDNA being barcoded/indexed by transposase-based tagmentation and pooled as per manufacturer's instructions. The prepared library was loaded on a FLO-MIN106 R9.4 flow cell (Oxford Nanopore Technologies, Oxford, UK) and sequenced using the MinION for 24 hours.

2.2.3.4 Data processing and analysis of Nanopore sequence data.

Raw FAST5 files were basecalled and de-multiplexed in real time, as reads were being generated, using Albacore v2.1 (Oxford Nanopore Technologies) into FASTQ files. Run metrics were generated using Nanoplot v1.8.1 ^[141] using default parameters. Reads were processed through Porechop v0.2.1 ^[142] using default parameters to remove any barcodes and adaptors used in SQK-RBK001. 96,788 reads (10,214,353 bases) and 430,911 reads (34,888,999 bases) were trimmed primarily making up the sequencing adapters for cases A and B respectively. 386 (0.39%) and 513 (0.11%) chimeric reads were split for cases A and B respectively. Trimmed FASTQs were speciated using Kraken v0.10.4 ^[164]. Krocus v1.0.0 ^[165] was used to assign a ST with the following parameters: --kmer 15, --min block size 300, and --margin 500 with the Achtman *E. coli* MLST scheme. *Stx* subtyping and serotyping was determined by aligning the basecalled reads using minimap2 v2.2 ^[146] and Samtools v1.1 ^[147] to a multifasta containing the *stx* and serotype encoding genes.

For reference-based variant calling FASTQ reads were mapped to the Sakai STEC O157:H7 reference chromosome (NC 002695.1) using minimap2 v2.2 ^[146]. VCFs were produced using GATK v2.6.5 ^[163] UnifiedGenotyper. Any variants called at positions that were within the known prophages in Sakai were masked from further analyses. To determine the optimum consensus cut-off for Nanopore variant detection the VCF was filtered with sequentially decreasing ad-ratio (consensus base sequence) values at 0.1 intervals. Using the Illumina variant calls as the gold standard, F1 scores (the weighted average of precision and recall) were calculated to determine the optimal ad-ratio for processing Nanopore data through GATK.

2.2.3.5 Comparison of Illumina and Nanopore discrepant SNPs.

Nanopolish v0.11.1 ^[145] was also used to detect methylation across the Nanopore data to compare to the discrepant positions. This was performed using the call-methylation function, searching for three types of methylation including the DNA adenine methyltransferase (Dam), DNA cytosine methylase (Dcm), and 5'-cytosine-phosphate-guanine-3' (CpG) models. The discrepant SNPs between the Illumina and Nanopore for both Case A and Case B were manually visualized in Tablet v1.17.08.17 ^[166] in order to elucidate the reason for the discrepancy. Discordant variants found within a homopolymeric region were also quantified.

2.2.3.6 Generation of Phylogenetic trees.

Illumina and Nanopore filtered VCFs for each sample were incorporated into SnapperDB v0.2.5 ^[94] containing variant calls from 4,471 other STEC CC11 genomes generated through routine surveillance by UKHSA. SnapperDB v0.2.5 ^[94]

was used to generate a whole-genome alignment of the 4,475 genomes (including both datasets for the selected strains for this study). Both methylated positions and prophage positions were masked from the alignment. The alignment was processed through Gubbins v2.0.0 ^[167] to account for recombination events. A maximum likelihood tree was then constructed using RAxML v8.1.17 (Stamatakis et al 2014 - 24451623).

2.3 Chapter 2 results.

2.3.1 Assembly-dependent results.

2.3.1.1 Initial draft assembly statistics and distance of true reference.

The draft assembly produced by Flye (pre-assembly correction) for strain Sakai assembled into two closed, circularised contigs, one chromosome (5,496,439bp), one plasmid (93,136bp). Comparing this to the publicly available reference (BA000007): chromosome 5,498,578bp and plasmid 92,721bp, revealed a discrepancy of 2,139bp on the chromosome and approximately 415bp on the plasmid.

Differences from the initial draft assembly and the publicly available reference genome included 7,468 single base substitutions, 17,176 and 501 single base insertions and deletions respectively, 642 and 22 double base insertions and deletions, 174 and 4 insertions and deletions between the size of 3bp to 50bp respectively and finally 1 deletion greater than 50bp (Tables 2.1 and 2.2).

2.3.1.2 Initial workflow results and selection of highest scoring workflow.

With only a single form of correction, Illumina reads led to a much more accurate result with both Pilon and Racon generating only 262 and 263 single base substitutions respectively when compared to using Nanopore reads as the correction dataset using four rounds of Racon or Nanopolish which generated approximately 5,300 and approximately 1,200 single base substitutions respectively (Figure 2.2).

The two most accurate workflows in terms of single base substitutions were, workflow 1; Nanopolish using Nanopore reads, one round of Pilon using Illumina

reads and one round of Racon using Illumina reads. This workflow generated 182 single base substitutions when using between 100x and 150x coverage of Nanopore reads (Figure 2.2, Table 2.1). The second workflow started with four rounds of Racon using Nanopore reads, Medaka with the STEC trained model using Nanopore reads followed by a single round of Pilon using Illumina reads and finally a single round of Racon using Illumina reads. This workflow generated 174 single base substitutions when using between 200x coverage of Nanopore reads (Figure 2.2, Table 2.2).

The only observable difference between using Nanopore reads from Guppy v3.2.6 and Guppy v5.0.11 was a much more accurate initial correction using racon in v5.011 (between 231-261 single nucleotide substitutions) compared to v3.2.6 (between 5,792-6,148 single nucleotide substitutions). Despite the initial correction the overall selection of workflow remained the same despite the version of Guppy used to basecall the Nanopore reads (Figure 22.).

Comparing the use of BWA MEM with default parameters vs BWA MEM -x intractg, produced a negligible result with a maximum difference in four substitutions observed through any workflow.

	Value	Substitutions	1bp Deletions	1bp Insertions	2bp Deletions	2bp Insertions	>2bp Deletions	>2bp Insertions
Guppy v3.2.6 Workflow 1 Default	Draft assembly	7468	501	17176	4	174	3	50
	Nanopolish default	1400	1187	4366	40	80	7	3
	Pilon default	194	53	115	2	2	2	0
	Racon default	188	28	94	0	2	2	0
Guppy v3.2.6 Workflow 1 Optimised	Draft assembly	7468	501	17176	4	174	3	50
	Nanopolish optimised	1335	1220	4260	48	79	6	3
	Pilon optimised	186	64	112	1	0	2	0
	Racon optimised	183	41	93	0	0	3	0
Guppy v5.0.11 Workflow 1 Optimised	Draft assembly	7468	501	17176	4	174	3	50
	Nanopolish optimised	947	771	3789	30	53	3	0
	Pilon optimised	175	40	83	1	0	1	0
	Racon optimised	173	23	76	0	0	1	0

Table 2.1 – Table showing discrepancies from both the default and optimised

parameters from workflow 1 using each Nanopore read set. Also showing the number of single base substitutions, single base indels, double base indels and indels greater than two base pairs.

	Value	Substitutions	1bp Deletions	1bp Insertions	2bp Deletions	2bp Insertions	>2bp Deletions	>2bp Insertions
Guppy v3.2.6 Workflow 2 Default	Draft assembly	7468	501	17176	4	174	3	50
	Racon x4 default	5562	5373	885	351	39	24	10
	Medaka trained	570	4960	362	288	9	14	1
	Pilon default	263	111	51	12	2	2	0
	Racon default	223	106	17	12	1	3	0
Guppy v3.2.6 Workflow 2 Optimised	Draft assembly	7468	501	17176	4	174	3	50
	Racon x4 optimised	5193	5332	726	347	40	20	20
	Medaka trained	431	4828	378	244	4	17	0
	Pilon optimised	200	84	53	10	2	2	0
	Racon optimised	174	84	24	10	1	2	0
Guppy v5.0.11 Workflow 2 Optimised	Draft assembly	7468	501	17176	4	174	3	50
	Racon x4 optimised	234	536	17	36	10	3	7
	Medaka trained	217	504	16	27	2	3	2
	Pilon optimised	200	17	6	1	0	1	0
	Racon optimised	175	16	6	0	0	1	0

Table 2.2 – Table showing discrepancies from both the default and optimised parameters from workflow 2 using each Nanopore read set. Also showing the number of single base substitutions, single base indels, double base indels and indels greater than two base pairs.

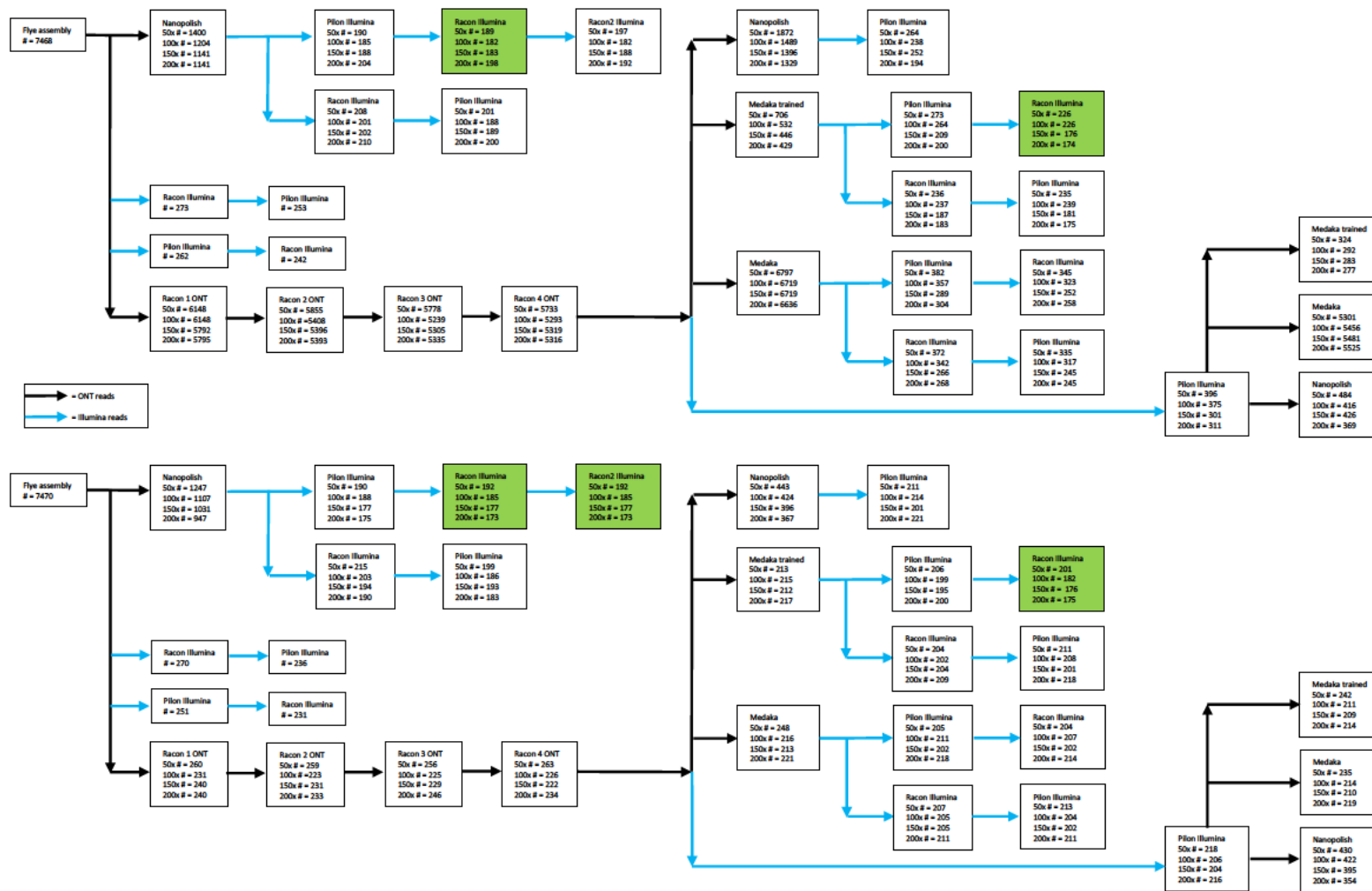


Figure 2.2 – Number of discrepant single base substitutions that remain for each correction method vs the publicly available reference genome (BA000007) using Guppy v3.2.6 (top) and Guppy v5.0.11 (bottom).

2.3.1.3 Final workflow optimisation.

Optimal workflow 1:

The optimisation of Nanopolish parameters of calling of variants led to the selection of $-d = 40$, a depth of 40 reads per position (default = 20) for the least number of single base substitutions being generated. The minimum support fraction had the lowest number of insertions and single base substitutions with the parameter set to $-m = 0.1$ (default = 0.2). For the Pilon parameters the `--mindepth` value of 0.05 (default = 0.1) showing the least number of differences with the number of discrepancies increasing as the `--mindepth` value increased. The `--minmq` that produced the least number of discrepancies was the same as the default (= 0) with discrepancies increasing as the value provided increased. The `--minqual` that produced the least number of discrepancies was 0 (default = 10). For the Racon parameters the `-q` value of 30 (default = 10) showing the least number of differences with the number of discrepancies increasing as the value provided increased. The default `-e` value of 0.3 was selected because all values provided did not affect the number of discrepancies.

Optimal workflow 2:

The optimisation of the four rounds of Racon parameters using Nanopore data, the `-q` value of 20 (default = 10) showing the least number of differences with the number of discrepancies increasing as the value provided increased. The default `-e` value of 0.1 was selected because all values provided increased the general number of insertions and deletions (regardless of size).

For both the Pilon and Racon (using Illumina reads) components of this workflow generated the same relative effect as workflow 1 with the same values being selected.

2.3.1.4 Final draft assemblies and description of remaining discrepancies.

The final optimised Workflow 1 including Nanopolish, Pilon and Racon with Nanopore reads from Guppy v3.2.6 generated a total discrepancy of 630 bases relative to the reference genome. Most of these discrepant positions were from two larger deletions one 143 bases and 162 bases in length (Table 2.1). This led to a total accuracy of 99.9887% (0.0113% error). Using Guppy v5.0.11, the finalised assembly generated a total discrepancy of 434 bases relative to the reference genome. Most of these discrepant positions were from a single large deletion of 162 bases in length (Table 2.1). This led to a total accuracy of 99.9922 % (0.0078% error).

The final optimised workflow 2 including four rounds of Racon, Medaka with a STEC trained model, Pilon and Racon with Nanopore reads from Guppy v3.2.6 generated a total discrepancy of 561 bases. Similar to the workflow above the majority of these positions were related to two large deletions, one being 95 bases and the other being 162 bases in length (Table 2.2). The deletion of 162 bases relative to the reference genome is the same size and position as the workflow above. This led to a total accuracy of 99.98985% (0.0102% error). Using Guppy v5.0.11, the final assembly generated a total discrepancy of 359 bases. Most of these discrepant

positions were from a single large deletion of 162 bases in length (Table 2.2). This led to a total accuracy of 99.9935% (0.0064% error).

For the most accurate workflow (#2 using Guppy v5.0.11 reads) there were 95.5% of discrepant bases (343/359) were located with prophage or prophage-like regions. 47.2% (162/343) of those found in the prophage and prophage-like regions were single base substitutions, the remaining 52.8% (181/343) were indel based discrepant bases (Figure 2.3).

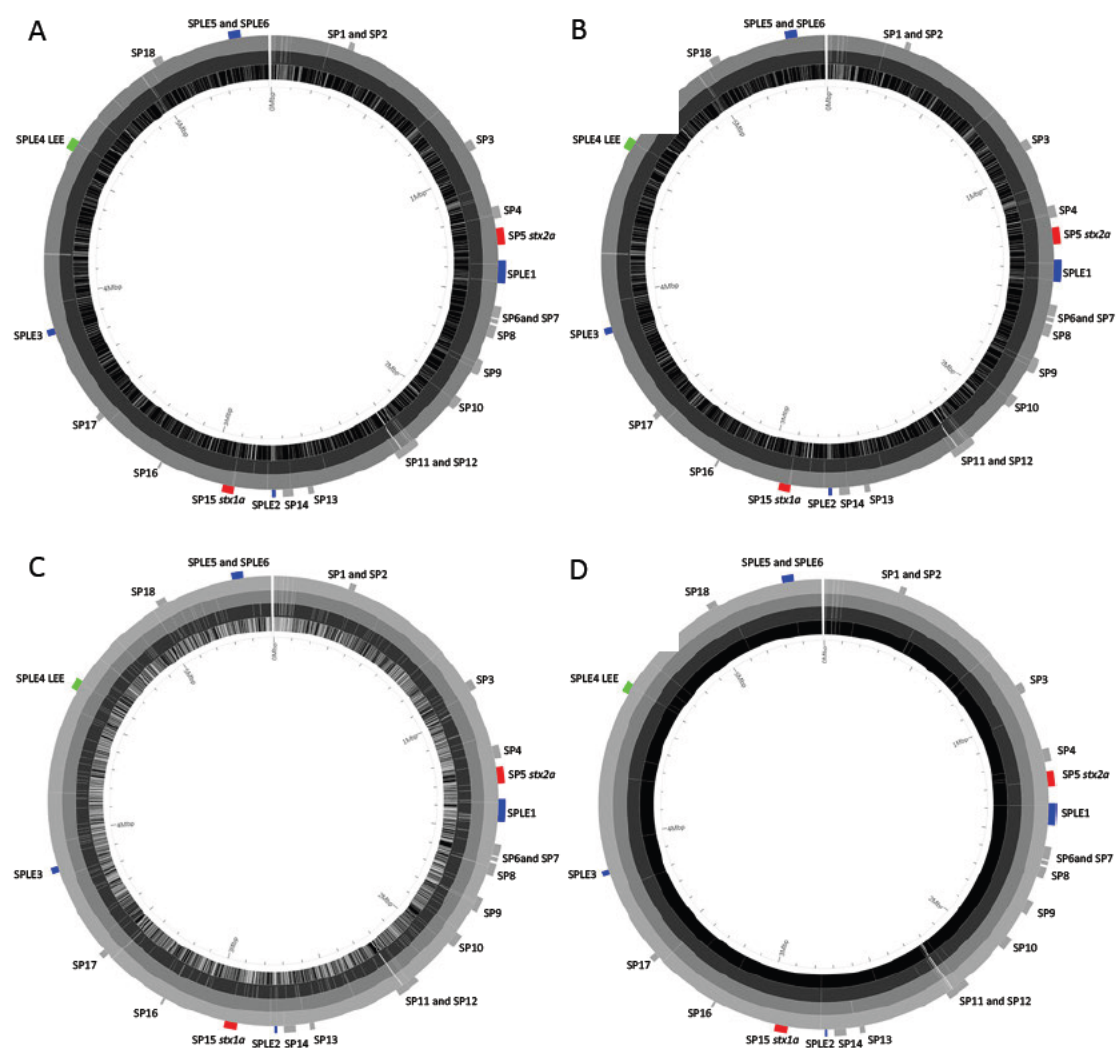


Figure 2.3 – Circos plot showing the number and position of single-base discrepancies (regardless of size) for the final draft assembly (chromosome) with Workflow 1 with Guppy v3.2.6 (A), Guppy v5.0.11 (B) and Workflow2 with Guppy v3.2.6 (C), Guppy v5.0.11 (D) when compared to BA000007.

Workflow 1 – Nanopolish [Nanopore reads] (inner black ring), Pilon [Illumina reads] (middle grey ring and Racon [Illumina reads] (outermost grey ring). Workflow 2 – 4x rounds of Racon [Nanopore reads] (inner black ring), Medaka STEC trained model for v3/v5 respectively [Nanopore reads] (second most inner grey ring), Pilon [Illumina reads] (second most outer grey ring and Racon [Illumina reads] (outermost grey ring).

Also showing on each diagram the relative positions of the non-*stx*-encoding prophages (Grey), prophage like regions (Blue), *stx*-encoding prophages (Red) and the Locus of Enterocyte Effacement (Green).

2.3.1.3 Comparison of prediction genome annotations.

Regardless of each genome annotation used both finalised assemblies from workflow 1 and workflow 2 agreed with the counts of rRNA and tRNA genes as the reference genome BA000007 whereas Illumina demonstrated a reduced count of each of the respective genes. A total CDS count across each corrected assembly varied when compared to BA000007 and for each genome annotation tool. For both Prokka Workflow 2 appeared to have slightly more CDS (n=9) called than BA000007 whereas workflow 1 had less (n=23) (Table 2.3). The same was noted when using Bakta (workflow 1 n=12, workflow 2 n=14). However, using PGAP, both workflows 1 and 2 had less CDS called than BA000007 (n=14, n=20 respectively). In each case

the Illumina-only assembly called much less CDS than the Nanopore assembled counterparts with the exception of PGAP which called the Illumina-only assembly with only 38 CDS less (Table 2.3).

Prokka	CDS	rRNA	tRNA
BA000007	5371	22	105
Workflow 1	5380	22	105
Workflow 2	5357	22	105
Illumina-only	5026	6	88
PGAP	Genes (total)	complete rRNAs	tRNA
BA000007	5445	8, 7, 7 (5S, 16S, 23S)	103
Workflow 1	5431	8, 7, 7 (5S, 16S, 23S)	103
Workflow 2	5425	8, 7, 7 (5S, 16S, 23S)	103
Illumina-only	5407	1, 1 (16S, 23S)	69
Bakta	CDS (predicted)	rRNA	tRNA
BA000007	5384	22	108
Workflow 1	5396	22	108
Workflow 2	5370	22	108
Illumina-only	5031	6	114

Table 2.3 – Table showing predicted CDS (number of coding sequences), rRNA (number of rRNA coding sequences) and tRNA (number of tRNA coding sequences) counts from genome annotation tools Prokka, PGAP and Bakta.

2.3.2 Assembly-independent results.

2.3.2.1 Optimisation of Nanopore variant calling.

To compare Illumina and Nanopore sequences within a standardised framework it was necessary to optimise the parameters for variant filtering using GATK2 to compensate for the lower per-read accuracy observed in the Nanopore data when compared to Illumina per-read accuracy. Case B was selected for the optimisation process, basecalls in the Nanopore data were classified as true-positive (variant base detected by both methods), false-positive (variant base in Nanopore, reference base in Illumina), true-negative (reference base in Illumina and Nanopore), or false-

negative results (variant base in Illumina, reference base in Nanopore). Prefiltering was performed by masking regions annotated as prophage in the reference genome and those that could not be accurately self-mapped with simulated reference Illumina FASTQ reads. This was to prevent regions where Nanopore reads could map to (and therefore identify variants) but were ambiguously mapped with Illumina reads from entering the comparison.

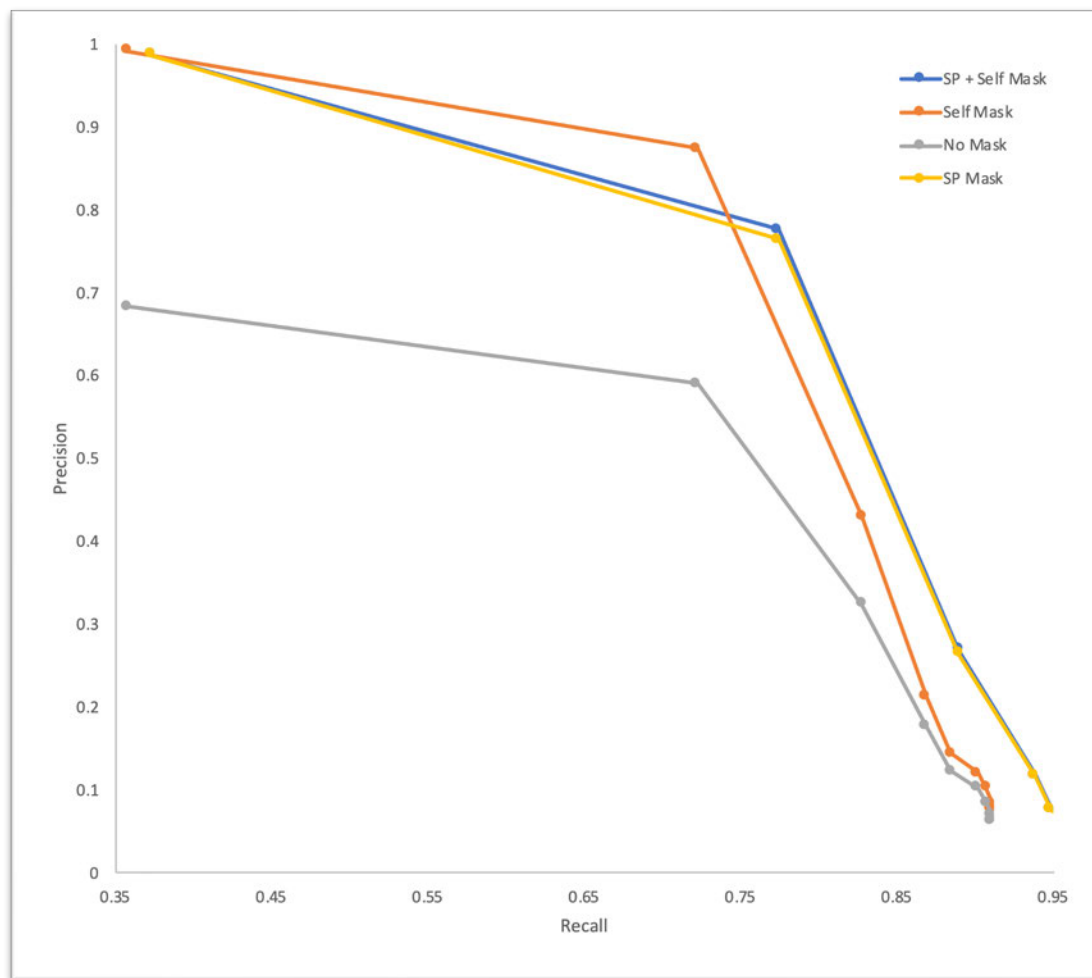


Figure 2.4 – Precision Vs Recall of variant calling (for case B) for an array of consensus ratio cut-offs and pre-masking strategies including masking positions annotated as ‘Sakai phage’ (‘SP’) and positions that are ambiguously self-mapped (‘Self’) with simulated Illumina FASTQs from the reference genome.

Figure 2.4 plots the precision (the proportion of true-positive results with respect to all positive calls) against the recall/sensitivity (the proportion of true-positive results identified with respect to all true-positive results) for an array of consensus ratio (ad ratio) cut-offs for each of the four masking strategies. Similar areas under the curve (AUCs) were achieved for the different masking strategies, with slightly higher precision at lower recall achieved with “self-masking” (AUC, 0.71) and slightly higher recall at lower precision with explicit masking of the Sakai prophage (AUC, 0.75). The absence of a masking strategy markedly affects the precision of variant calling with Nanopore data, in comparison with Illumina as gold standard (AUC, 0.30). To identify the optimum consensus cut-off for filtering Nanopore variants processed through GATK the F1 score was calculated at each consensus cut-off. A consensus cut-off of 0.8 maximized the precision and recall (Figure 2.5) irrespective of the filtering methods.

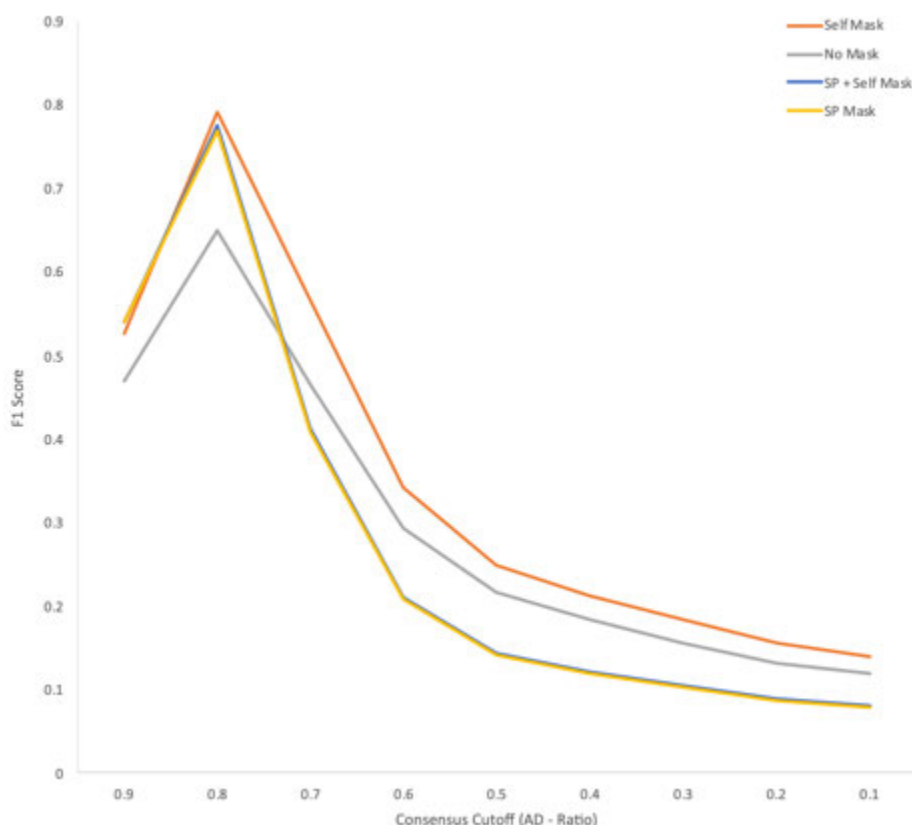


Figure 2.5 – F1 Score for an array of consensus ratio cut-offs (0.1 through to 0.9 at 0.1 increments) and pre-masking strategies including masking positions annotated as ‘Sakai phage’ (‘SP’) and positions that are ambiguously self-mapped (‘Self’) with simulated Illumina FASTQs from the reference genome.

2.3.2.2 Investigation of the discrepant variants identified between Illumina and Nanopore.

After optimised quality and prophage filtering for Nanopore and Illumina sequencing data there were 266 and 101 discordant base positions for Cases A and B, respectively. The majority of discrepancies were where the Nanopore data identified a variant position which was not identified in the Illumina data (261 of 266 [98.1%] and 95 of 101 [94.1%] discrepant base positions for Cases A and B, respectively). In

contrast the Illumina data identified five (1.9%) discrepant base positions as variants for Case A and six (5.9%) for case B (Table 2.4) not identified by the Nanopore data. For both cases the most common discrepant variants were adenines misclassified as guanines (A → G) in the Nanopore data with respect to the Illumina data (and reference), accounting for 68.0% (181/266) for Case A and 72.3% (73/101) for Case B (Table 2.5). The second most common discrepancy was thymine being classified as cytosine (T → C) in the Nanopore data accounting for 29.7% (79/266) in Case A and 20.8% (21/101) in Case B (Table 2.4). Of the transitions described above, 97.7% (Case A) and 93.1% (Case B) occurred when the variant was between 2 homopolymeric regions of either multiple cytosines and guanines (Figure 2.5). These homopolymeric regions were similar to described DNA cytosine methylase (Dcm) binding sequences (Gomez-Eichelmann et al 1991). Nanopolish was subsequently used to identify likely Dcm, 5'-cytosine-phosphate-guanine-3' (CpG), and DNA adenine methyltransferase (Dam) methylation sites in the Nanopore sequencing data and confirmed that 260 of 266 (97.7%) and 94 of 101 (93.1%) discrepant variants in the Nanopore data were classed as methylated for Cases A and B, respectively. All of these were determined to be Dcm methylation for both cases.

Once the methylated positions were masked from the analysis, there were a total of 6 (5 discrepant variants in Illumina and 1 Nanopore) and 7 (6 discrepant variants in Illumina and 1 Nanopore) discrepant SNPs between the Nanopore and Illumina data, for Cases A and B, respectively (Tables 2.4 and 2.6). Four discrepant Illumina variants are shared by both Case A and Case B. One shared variant was found in a non-coding region; another shared variant was found in *rhcC* encoding an RHS (rearrangement hotspot) protein defined by the presence of extended repeat regions.

Two further shared variants were found in *dadX*, an alanine racemase gene. *dadX* is a paralogue of *alr*, also annotated as an alanine racemase in the Sakai reference genome with high nucleotide similarity (>75% nucleotide identity). Both intra- and inter-gene repeats are known to be regions of potential false-positive calls with Illumina data due to ambiguous mapping. Of the 7 variants in the Illumina data found in either or both Case A and B, 5 were found to be homoplastic in the O157 population of 4,475 Illumina sequences, arising independently, multiple times.

Variants and reason for omission.	Case A		Case B	
Total # of variants against the reference genome post quality filtering.	2076		1424	
Total # of variants with masked due to location in phage	708		531	
Total # of discrepant variants called between case A and B alone.	266		101	
Variants and reason for omission.	Illumina VCF	ONT VCF	Illumina VCF	ONT VCF
# of discrepant variants in each VCF.	5	261	6	95
# of discrepant variants with methylated positions masked.	0	260	0	94
Final discrepant variants.	5	1	6	1

Table 2.4 – Table showing the breakdown of the total number of variants of each technology against the reference genome, followed by the numbers of masked variants within prophage or methylated positions.

	Position					Case	
	-2	-1	Variant	+1	+2	A	B
Reference	C	C	A	G	G	A > G Transition	69.62% (n=181)
Alignment	C	C	G	G	G		
Reference	C	C	T	G	G	T > C Transition	30.38% (n=79)
Alignment	C	C	C	G	G		
Total						100% (n=260)	100% (n=94)

Table 2.5 – Table showing the two most common discrepancies in the Nanopore optimised GATK variant calls and a breakdown of the relative proportions of these transitions compared to the total number of discrepant SNPs called for both cases.

SNP	Position	BASE in Ref	BASE in Illum	Depth in Illum	BASE in ONT	Depth in ONT	Variant	Locus tag	Annotation
1	270,595	C	A	19	C	207	A	ECs0237	<i>rhsC</i>
2	379,516	A	G	52	A	124	G	NON CODING	-
3	1,681,338	C	G	44	C	86	G	ECs1685	alanine racemase 2
4	1,681,339	G	C	41	G	86	C	ECs1685	alanine racemase 2
5	2,033,176	T	G	34	T	85	G	ECs2049	hypothetical protein
6	2,731,621	A	C	52	A	73	C	NON CODING	-
7	4,901,209	A	A	49	G	102	G	ECs4834	superoxide dismutase <i>sodA</i>

Table 2.6 – Table showing the final discrepant SNPs (n=7) between the Illumina (Illum) data and Nanopore (ONT) data for case B. Also shown is the reference base, the Illumina variant called base and read depth at that position and the same for the Nanopore data. Finally, also included is the locus tag relative to the reference genome and the associated gene annotation.

2.3.2.3 Phylogenetic analyses.

Using the optimised variant calling parameters both case A and B clustered phylogenetically in sub-lineage 1c within a dense reference database of STEC O157:H7 genomes ($n = 4,475$). However, the genomes were located in distinct subclades (Figure 2.6). It was, therefore, unlikely that the isolates originated from the same source, and it was concluded that Cases A and B were not epidemiologically linked. Following phylogenetic analysis of the Illumina SNP typing data (Figure 2.6), Case A was designated a sporadic case. However, Case B clustered with a concurrent outbreak, already under investigation, comprising three additional cases. The Illumina sequence linked to Case B was zero SNPs different from the other three cases in the cluster, whereas the Nanopore sequence was 7 SNPs different, when excluding the methylated positions (Table 2.6). Based on the Nanopore sequencing data alone, this discrepancy would have led to uncertainty as to whether Case B was linked to the outbreak.

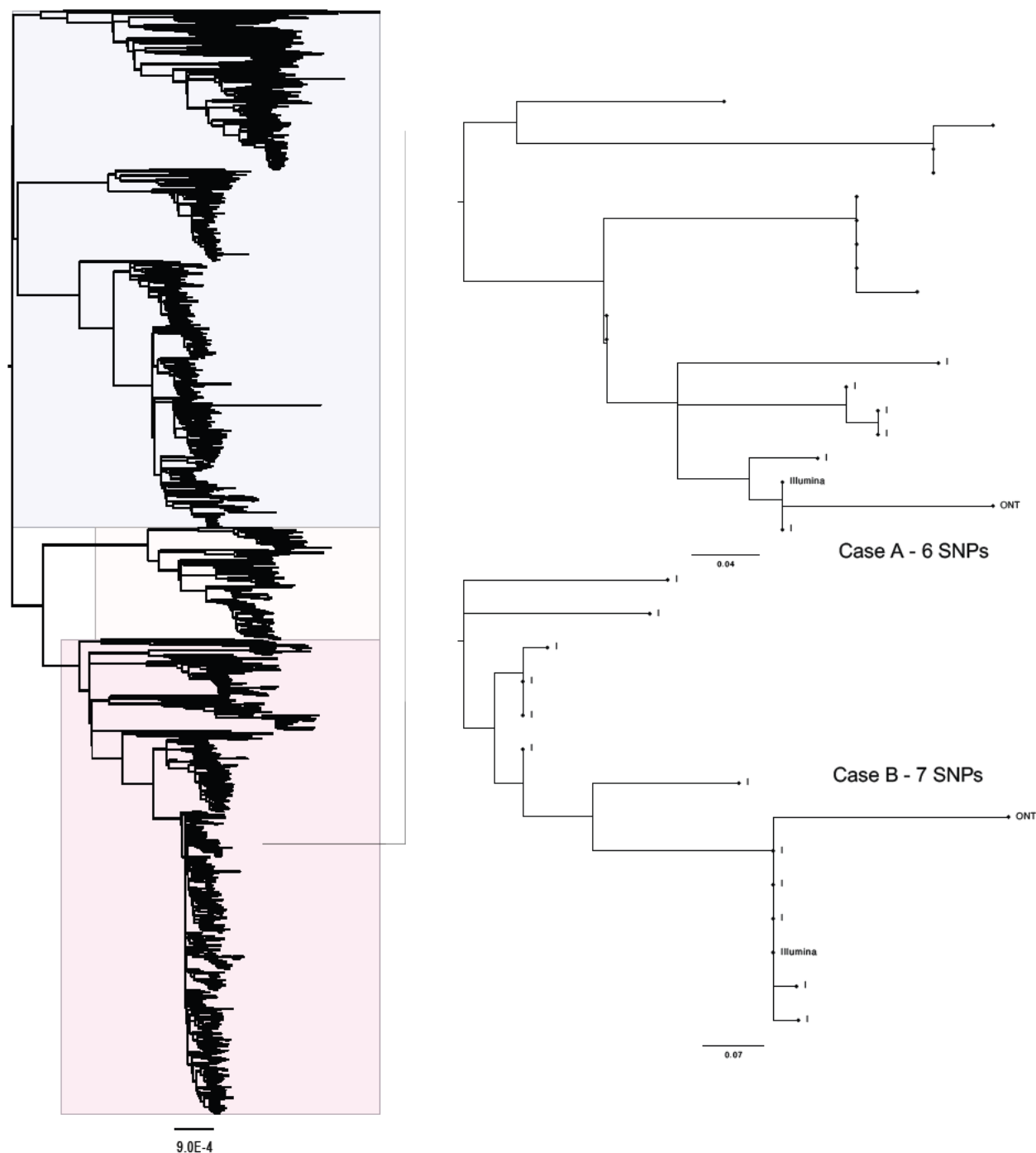


Figure 2.6 – Maximum likelihood tree, of a soft core alignment (Includes variants where 80% of cases share said variant) of 4,475 genomes showing the tree lineages (I [pink], I/II [yellow] and II [blue]) of STEC O157:H7 (Clonal Complex 11). Also showing where Nanopore and Illumina sequencing data is placed within the tree for each of the two cases. All methylated positions and prophage regions have been

masked. Values represent the SNP differences between the Illumina and Nanopore data for both cases.

2.4 Chapter 2 discussion.

2.4.1 Assembly-based correction workflow optimisation.

Comparing the tools that utilised short reads to correct draft assemblies, Pilon followed by Racon typically generated a better result than Racon followed by Pilon in terms of single base substitutions. Nanopolish outperformed Medaka when run on default parameters, however, once a trained model can be applied to Medaka, it then outperforms Nanopolish. Therefore, Nanopolish is the preferred option in the absence of a trained model to correct the draft assembly. Comparing all the different workflows (Figure 2.2) tested within this study, it was clear that a more accurate workflow included the use of short reads. The remaining error within the Nanopore reads still results in a more erroneous sequence post-correction when compared to Illumina [169].

An optimised workflow of Racon (four rounds with Nanopore reads), a STEC trained Medaka model (with Nanopore reads), a single round of Pilon (with Illumina reads) and a single round of Racon (with Illumina reads) produced the closest draft assembly compared to the Sakai O157:H7 reference genome with a final accuracy of 99.9935% (0.0064% error) which equates to about 200 bases difference. A second workflow, using Nanopolish (with Nanopore reads) a single round of Pilon (with Illumina reads) and a single round of Racon (with Illumina reads) was an acceptable alternative with a final accuracy of 99.9885% (error 0.0115%), particularly if no trained models are available. Regardless of which workflow was used, the more recently released Guppy v5 produced a more accurate result than Guppy v3.

2.4.2 Assembly-based correction discrepancies related to the STEC O157:H7 genome.

Despite the optimisation process there was still a large proportion of discrepant SNPs appearing in homologous prophage regions of the STEC genome (Figure 2.3). STEC typically harbour 16-19 prophages [75,81,111,170] and the Sakai reference genome has 18 prophage and 6 prophage-like regions [75]. The homologous sequences in these prophage regions lead to ambiguity when aligning short reads as the read length does not span homologues sequenced thus failing to incorporate the unique sequence flanking the homologues [171]. Ambiguously aligned reads to the prophage regions was the most likely reason for the 95.5% of false positive and false negative substitutions and indels. (Figure 2.3).

2.4.3 Assembly-independent methodology to determine STEC genome relatedness.

The two isolates were sequenced using Nanopore were identified as STEC O157:H7 ST11 *stx2a/stx2c* and it was possible to calculate the genetic distance between them and to a large database of 4,475 STEC O157:H7 Illumina derived genomes. This demonstrates that Nanopore sequencing technologies can be integrated into the current UKHSA STEC O157:H7 workflows. However, we have highlighted some optimisation must take place in order for this comparison to work, primarily accounting for the lower per-read accuracy of Nanopore reads and secondly accounting for the methylated positions which the tested basecallers have struggled to determine (Table 2.5).

2.4.4 Assembly-independent methodology limitations and discrepancies.

A current limitation of Nanopore sequencing is its lower per-read accuracy when compared to short-read technologies [172,173,174,175,176]. This accuracy has improved as the technology has matured but still falls short of the 99% per-read accuracy offered by short-read platforms [175]. There are several factors that contribute to the current read accuracy in the nanopore data including structural similarity of nucleotides, simultaneous influence of multiple nucleotides on the signal, the non-uniform speed at which nucleotides pass through the pore, and the fact that the signal does not change within homopolymers [175].

Although analysis of the Illumina and Nanopore sequencing data placed the sequences on the same branch on the phylogeny, there were six SNP discrepancies between the sequences generated by the two different workflows, even after optimisation of the parameters. The vast majority of the discrepant SNPs (261/266 [98.1%] and 95/101 [94.1%] for Cases A and B, respectively) were attributed to variants identified in the Nanopore data and not the Illumina data. The majority of discrepancies (97.7% in Case A and 93.1% in Case B) were found in sequences that are the same as the known 5-methylcytosine motif sequences, CC(A/T)GG in the Nanopore data [177,178]. Following a search of the Nanopore discrepant SNPs for CpG, Dam, and Dcm methylation using Nanopolish, the majority (97.7% and 93.1% for case A and B, respectively) of the Nanopore discrepant SNPs were identified in Dcm (5-methylcytosine) methylated regions.

As Nanopolish is detecting these methylated positions with the use of the raw FAST5 data, it is suggested that these particular discrepancies appear during the base-

calling process. Albacore, the basecaller, handles most methylation well across the three methylation models searched for by Nanopolish, e.g., only 94 of 13,504 methylated positions were considered to be incorrect by base calling for Case B. However, for mapping-based SNP typing, this level of error in base calling means that it is not possible to accurately determine the number of SNPs, thus potentially obscuring the true phylogenetic relationship between isolates of STEC O157:H7. The optimisation of variant filtering was performed using the Illumina data as a gold standard. However, it is likely that the alignment of the Illumina data might have generated false SNPs based on reads mapping to ambiguous regions of the genome, whereas the long reads obtained using the Nanopore workflow are able to resolve these ambiguous regions and call variants, or not, at these positions correctly. Because the Illumina data were used as the gold standard, in this scenario SNPs produced in the Illumina data would have been classed incorrectly as false-negative results in the Nanopore data. Discrepant variants identified in the Illumina data were attributed mainly to false mapping/aligning of Illumina reads to homologous regions of the reference genome, variants that were misidentified at the same position independently in Cases A and B.

2.4.5 Nanopore technological advancements.

With the recent development of both new more accurate chemistries being produced by Oxford Nanopore Technologies and the development of new software will decrease the per-read error rates of Nanopore reads, in particular the development of duplex and ligation kit 14. This will produce more accurate reads for any assembly-independent analyses and also more accurate genome assemblies with

reduced single base and indels in the initial draft. This creates the possibility of being able to compare full *de novo* assemblies in a pairwise format for variant detection/calling to determine relatedness of STEC and other bacterial genomes. Rang *et al* 2018 ^[175], stated how the academic and scientific community can make valuable contributions to improving Nanopore read accuracy by systematically comparing computational strategies as highlighted in this project and elsewhere ^[169].

2.5 Chapter 2 conclusions.

Currently, both short- and long-read technologies are used for public health surveillance, and with more long-read and hybrid *de novo* assemblies being produced from public health settings it is important to benchmark and understand how discrepancies are occurring within these assemblies. The limitations of correcting bacterial genomes with a lot of homologous sequences such as STEC O157:H7 must be considered as the introduction of errors is common within these regions. This concept is very important when performing downstream analyses that may impact on public health actions.

Similarly, it was demonstrated that Nanopore sequencing can be used to rapidly rule out an epidemiological link between two cases admitted to the same hospital on the same day with symptoms of HUS using assembly-independent optimised bioinformatics methodology.

The per-read error rate exhibited by Nanopore sequencing workflows continues to improve because of developments in the pore design, the library preparation methods, innovations in basecalling algorithms, and the development of post-sequencing correction tools, such as Nanopolish and Medaka. Currently, both short-

and long-read technologies are used for public health surveillance, and there is a need to integrate the outputs so that all data can be analysed in the same way. With ongoing updates to the chemistry and software tools will facilitate the robust variant calling and assembly correction enabling Nanopore to compete with short-read platforms, ultimately enabling the two technologies to be used interchangeably in clinical and public health settings.

Chapter 3:

Detection and characterisation of Shiga toxin-encoding prophages within highly pathogenic STEC O157:H7 UK sub-lineages.

3.1 Chapter 3 background.

There are three main lineages of STEC O157:H7 (I, II and I/II) and eight sub-lineages (Ia, Ib, Ic, IIa, IIb, IIc, I/IIa and I/IIb). In the UK, the outbreaks of STEC O157:H7 in the 1980s were typically caused by strains belonging to sub-lineage Ia (mainly comprising PT1 and PT4), sub-lineage I/IIa (comprising PT2) and sub-lineage I/IIb (comprising PT49) ^[100]. Throughout the 1990s, these three lineages declined and almost disappeared. Concurrently, there was a sharp rise of sub-lineage Ic (mainly comprising PT21/28), in addition to a steady annual increase in the number of cases of sub-lineage IIc (mainly comprising PT8) ^[80,100]. Since 2012, the number of cases of sub-lineage Ic PT21/28 has declined and an uncommon PT8 variant belonging to sub-lineage IIb has emerged ^[103].

All the dominant UK sub-lineages over time encoded *stx2a*, with the exception of sub-lineage IIc (PT8), which is not associated with HUS cases in the UK ^[80]. The acquisition of a *stx2a*-encoding bacteriophage appears to have coincided with a rapid emergence and subsequent dominance of each sub-lineage in human clinical infections ^[80,103].

In this chapter, I describe the use of Oxford Nanopore platforms to sequence representative strains of STEC O157:H7 from each of the UK sub-lineages. The aim was to characterise and compare the *stx*-encoding prophage to determine the similarity of the *stx*-encoding prophage acquired by each sub-lineage. Understanding the nature of the *stx*-encoding bacteriophages, will contribute to our understanding of the emergence of highly pathogenic strains of STEC O157:H7.

3.2 Chapter 3 methods.

3.2.1 Strain selection and bacterial culture.

Six strains of STEC O157:H7 were selected for sequencing from the UKHSA STEC archive based on being the earliest recorded representative of each of the UK domestic sub-lineages that acquired the *stx2a*-encoding prophage (Table 3.1, Figure 3.1). A further non-domestic associated sample was included associated with foodstuffs import to the UK (Table 3.1).

To add context to the analyses, eleven publicly available sequences were included. Seven of the publicly available samples originated from the UK, of which five were the cause of four previously published outbreaks [68,139,179], three were from the USA [16,180] and single genome from Japan [181] (Table 3.1, Figure 3.1).

All strains selected were cultured from long-term storage at UK Health Security Agency (UKHSA) [known previously as Public Health England (PHE)] using 4ml of double-strength nutrient broth (Oxoid) and incubated at 37°C overnight.

3.2.2 Illumina sequencing and data processing.

Genomic DNA was extracted from cultures of STEC O157:H7 using the QIAAsymphony system (Qiagen) as part of the UKHSA routine service. Whole genome sequencing was performed on the HiSeq 2500 platform (Illumina) with the fast protocol, following library preparation using the Nextera XP kit (Illumina). FASTQ reads were processed using Trimmomatic v0.27 [140] to remove bases with a PHRED score of <30 from the leading and trailing ends, with reads <50 bp after quality trimming discarded.

3.2.3 Nanopore sequencing and data processing.

Qiagen genomic tips, midi 100/G (Qiagen), was used to extract genomic DNA for samples going for Nanopore sequencing with minor alterations including no vigorous mixing steps (mixing performed by inversion instead) and elution into 100µl double processed nuclease-free water (Sigma-Aldrich). Genomic DNA for each sample was quantified using the HS (high sensitivity) dsDNA assay kit (Thermofisher Scientific), following the manufacturer's instructions and the quantification performed on the Qubit. Library preparation was performed for several instances using both rapid barcoding (SQK-RBK001 and SQK-RBK004) and native barcoding kits (SQK-LSK108 and EXP-NBD103) (Oxford Nanopore Technologies). The prepared sequencing libraries were loaded onto FLO-MIN106 R9.4.1 flow cells (Oxford Nanopore Technologies) and sequenced using the MinION (Oxford Nanopore Technologies) for 48 hours.

Basecalled and de-multiplexing on raw FAST5s was performed using Albacore v2.3.3 (Oxford Nanopore Technologies) into FASTQ format. De-multiplexing was performed for a second iteration using Deepbiner v0.2.0 ^[182]. Sequencing metrics were calculated using Nanoplot v1.8.1 ^[141]. Reads were trimmed to remove the y-adapter and the barcode adapter (index adapter), and chimeric reads split using Porechop v0.2.4 ^[142]. Finally, the trimmed reads were filtered using Filtlong v0.1.1 ^[143] with the following parameters, min_length=1000 bp, keep_percent=90 and target_bases=550 Mbp, to generate approximately 100× coverage of the STEC genome with the longest and highest-quality reads.

3.2.4 *De novo* assembly, correction, and re-orientation.

A three-method approach was used to generate the best assembly of each sample that underwent Nanopore sequencing. Firstly, Canu v1.7 ^[183] was used with default parameters with the trimmed nanopore FASTQ files. Secondly, each of the filtered Nanopore FASTQ files were assembled using Unicycler v0.4.2 ^[169], with the following parameters min_fasta_length=1000 bp, mode=normal. Finally, Flye v2.4.2 ^[144], using default parameters. Where the assembly N50 was at its highest, number of contigs at its lowest and with the total assembly size within 5.3–6.0 Mbp, that iteration was selected and taken forward.

Assembly correction to produce the most accurate assembly possible was performed in a three-step process. Firstly, polishing began with Nanopolish v0.11.1 ^[145] using both the trimmed nanopore FASTQ and FAST5 files for each respective sample accounting for methylation using the --methylation-aware=dcm and --min-candidate-frequency=0.5 with alignments generated from Minimap2 v2.7 ^[146] and Samtools v1.7 ^[147]. Secondly, Pilon v1.22 ^[148] using the equivalent Illumina equivalent FASTQ reads and an alignment of the Illumina FASTQs to the draft assembly produced by Nanopolish as the query dataset with the use of BWA-MEM v0.7.17 ^[149] and Samtools v1.7 ^[147]. The final step involved, Racon v1.2.1 ^[150] also using BWA-MEM v0.7.17 ^[149] and Samtools v1.7 ^[147] was used with the Illumina reads for two iterations/cycles to produce a final assembly for each of the samples.

As the chromosome from each assembly was circularised and closed, they were re-orientated to start at the *dnaA* gene (GenBank accession no. NC_000913) from *E. coli* K12, using the --fixstart parameter in Circlator v1.5.5 ^[184]. A personalised database (an amino acid multi-FASTA that included all genes annotated in the

publicly available samples used in this study) was used to annotate the final assemblies with Prokka v1.13 [155].

3.2.5 Prophage detection and comparisons.

All chromosomal prophages across all samples were detected and extracted using the updated Phage Search Tool (PHASTER) [185]. Regardless of prophage size or PHASTER quality score, and any detected prophages separated by less than 4000 bp were conjoined into a single phage and all extracted using Propi v0.9.0, as described from Shaaban *et al* 2016 [81]. Prophages were trimmed to remove any non-prophage genes and were re-annotated using Prokka v 1.13 [155] with the same a personalised database described in Section 3.4.2.

3.2.6 Mash and *stx*-encoding prophage phylogeny.

For all *stx*-encoding prophages found in the publicly available STEC genomes in Table 3.1, Mash v2.2 [186] was used to sketch (sketch length 1000 bp, kmer length 21) the extracted prophages. Jaccard distances were calculated between the prophages in a pairwise format and a neighbour-joining tree constructed and visualised using FigTree v1.4.4 [187].

3.2.7 Visualisation tools and phylogenetic context.

To provide context for the 18 nanopore sequenced (n=7) and publicly available (n=11) samples, a maximum-likelihood phylogeny was recreated combining samples analysed in this study with 88 Illumina generated genomes (105 in total) from the UKHSA STEC collection (CC11 specifically). Each of the 88 selected genomes is

from a unique 250 single-linkage hierarchical cluster that was calculated using SnapperDB v0.2.6 ^[94]. SnapperDB v0.2.6 ^[94] was also used to generate a whole-genome alignment of all 105 genomes which was subsequently processed through Gubbins v2.0.0 ^[167] to identify and mask any recombinant or horizontally acquired sequences. The phylogeny was created by RAxML v8.2.8 ^[168]. Finally, all maximum-likelihood phylogenetic trees were annotated and visualised using FigTree v1.4.4 ^[187]. All gene-based diagrams were constructed using Easyfig v2.2.3 ^[188].

3.3 Chapter 3 results.

3.3.1 Genomic features of Nanopore sequenced samples.

The seven isolates, selected for Nanopore sequencing from the UKHSA [previously PHE] archive generated complete assembled, closed chromosomes with one or more plasmids. The sub-lineage Ia PT4 (E30228) genome and sub-lineage IIb PT8 (315176) genome each assembled into a single chromosome (5,416,109bp and 5,579,120bp respectively) and two plasmids (Table 3.1). The remaining four isolates assembled into a single complete closed chromosome of between 5,359,964bp and 5,571,891bp in size and one plasmid (Table 3.1). The non-domestic strain assembled into a chromosome of 5,753,376bp and a single plasmid (Table 3.1). The pO157 (IncFIB) plasmid was found in all samples sequenced in this study. Prophages counts varied between the seven isolates and thus sub-lineages of STEC O157:H7. They ranged from 14 prophages in the isolate belonging to sub-lineage I/IIa PT2 to 19 prophages in the non-domestic PT24 strain sequenced (Table 3.1, Figure 3.2). Finally, the *stx* subtypes associated with each strain were either *stx1a* and/or *stx2a* and/or *stx2c* (Table 3.1).

Strain ID	CC11 Sub-lineage	Phage Type	Stx-profile	Chromosome size (bp)	No of prophages	Stx1a prophage SBI and size (bp)	Stx2a prophage SBI and size (bp)	Stx2c prophage SBI and size (bp)
UKHSA archive genomes								
E30228	Ia	PT4	<i>Stx1a/2a</i>	5,416,109	15	<i>yehV</i> (47,594)	<i>wrbA</i> (62,890)	-
E34500	I/IIa	PT2	<i>Stx2a/2c</i>	5,359,964	14	-	<i>argW</i> (62,149)	<i>sbcB</i> (57,463)
E45000	I/IIb	PT49	<i>Stx2a</i>	5,386,698	17	-	<i>sbcB</i> (44,014)	-
E116508	Ic	PT21/28	<i>Stx2a/2c</i>	5,571,891	17	-	<i>argW</i> (71,870)	<i>sbcB</i> (59,105)
315176	IIb	PT8	<i>Stx2a</i>	5,579,120	16	-	<i>sbcB</i> (61,851)	-
267849	IIa	PT34	<i>Stx2a/2c</i>	5,510,912	16	-	<i>yecE</i> (47,242)	<i>sbcB</i> (61,840)
194195	IIa	PT24	<i>Stx2a/2c</i>	5,753,376	19	-	<i>argW</i> (74,662)	<i>sbcB</i> (60,586)
Publicly available genomes								
9000	Ic	PT21/28	<i>Stx2a/2c</i>	5,516,497	17	-	<i>argW</i> (65,158)	<i>sbcB</i> (57,408)
397404	Ic	PT21/28	<i>Stx2a/2c</i>	5,618,435	13	-	<i>argW</i> (70,472)	<i>sbcB</i> (59,098)
155	Ic	PT32	<i>Stx2a</i>	5,513,008	18	-	<i>yecE</i> (50,015)	-
350	IIc	PT8	<i>Stx1a/2c</i>	5,411,823	16	<i>yehV</i> (49,867)	-	<i>sbcB</i> (57,747)
272	I/IIa	PT2	<i>Stx2a</i>	5,474,193	16	-	<i>argW</i> (65,675)	-
644	IIc	PT8	<i>Stx1a/1a/2c</i>	5,831,209	19	<i>yehV</i> (49,544) <i>argW</i> (64,569)	-	<i>sbcB</i> (58,210)
180	IIc	PT54	<i>Stx1a/1a/2c</i>	5,509,528	15	<i>yehV</i> (49,544) <i>argW</i> (64,569)	-	<i>sbcB</i> (61,558)
Sakai	Ia	-	<i>Stx1a/2a</i>	5,498,450	18	<i>yehV</i> (47,650)	<i>wrbA</i> (62,142)	-
EDL933	Ia	-	<i>Stx1a/2a</i>	5,547,323	14	<i>yehV</i> (47,596)	<i>wrbA</i> (61,066)	-
EC4115	I/IIa	-	<i>Stx2a/2c</i>	5,572,075	17	-	<i>argW</i> (71,540)	<i>sbcB</i> (60,476)
TW14359	I/IIa	-	<i>Stx2a/2c</i>	5,528,136	17	-	<i>argW</i> (71,540)	<i>sbcB</i> (60,476)

Table 3.1 – Summary of the UKHSA archived strains and publicly available strains used within this study, detailing their strain ID, lineage, phage type (PT), *stx*-subtype, chromosome size, number of prophages detected in the chromosome and *stx*-encoding prophages with site of Shiga-toxin producing bacteriophage integration (SBI) and size (bp).

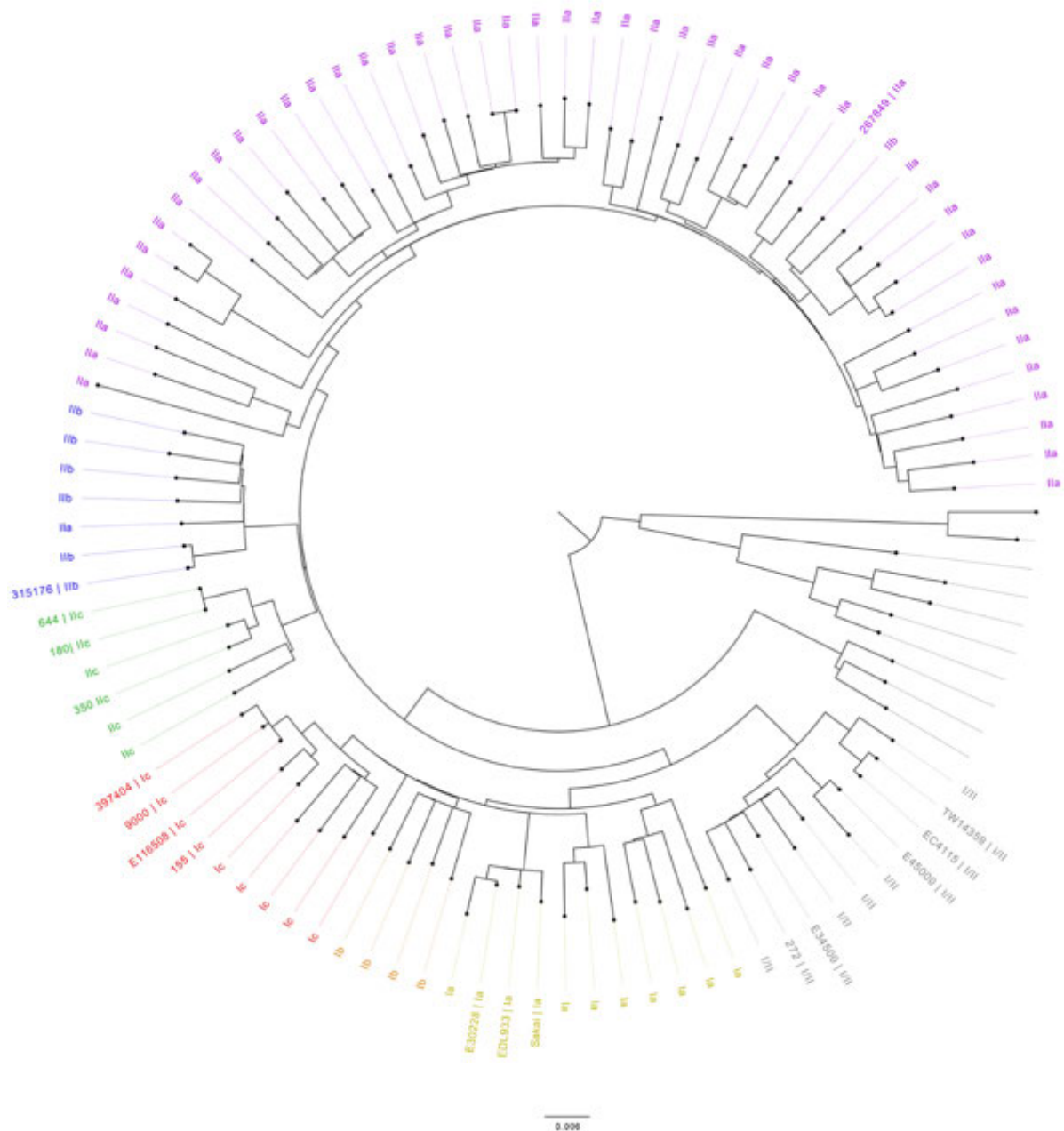


Figure 3.1 – Maximum-likelihood phylogenetic tree of 105 CC11 STEC O157:H7 genomes including n=17 (labelled) publicly available genomes and domestic UK nanopore sequenced genomes produced during this project. Sub-lineages are coloured as follows: I/II, grey; Ia, yellow; Ib, orange; Ic, red; IIa, purple; IIb, blue; IIc, green. Scale bar indicates substitutions per site of a soft-core alignment of variants from each sample in the phylogeny.

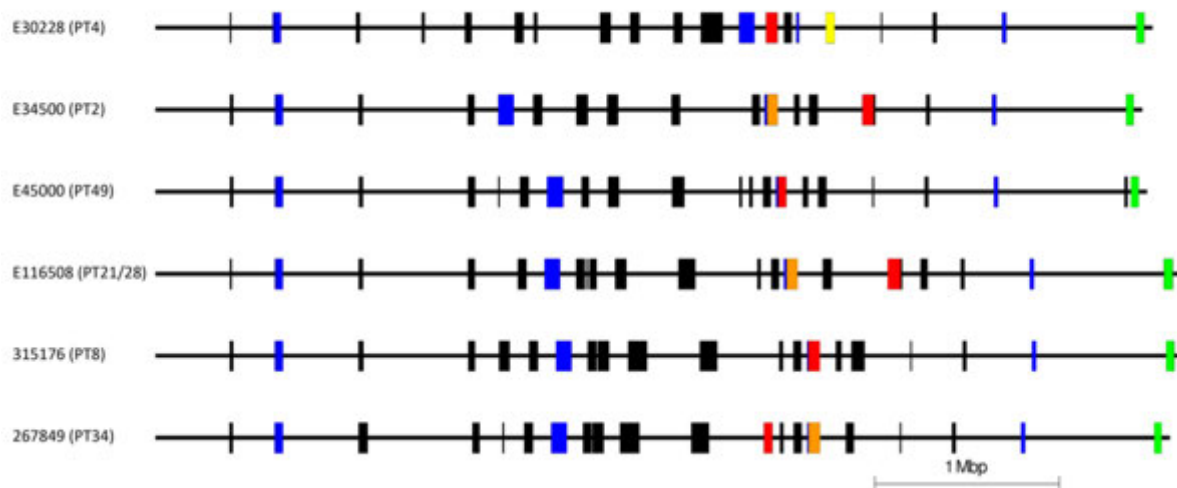


Figure 3.2 – Easyfig plot showing a pairwise alignment of the chromosome (horizontal lines) and prophage content (blocks) within the samples sequenced in this study (in descending order PT4, PT2, PT49, PT21/28, PT8 and PT34). *stx2a*-encoding, *stx2c*-encoding and *stx1*-encoding prophages are blocks highlighted in red, orange and yellow, respectively. Prophage-like elements are blocks coloured blue, and the locus of enterocyte effacement are blocks shown in green. Finally, the non-*stx*-encoding prophages are blocks coloured black. Scale equates to 1 million base pairs.

3.3.2 Comparison of the *stx1a*-encoding prophages.

Of the six of the domestic isolates analysed in this project that contained a *stx1a*-encoding prophage encoding, one was sequenced during this project and the remaining five were publicly available genome. (Table 3.1, Figure 3.3 and Figure 3.4). One of the first STEC O157:H7 isolated in 1983 belonged to sub-lineage Ia, PT4 (E30228) harboured a *stx1a*-encoding prophage which shared high sequence similarity to both the *stx1a*-encoding prophage from two international publicly

available genomes, strain Sakai and EDL933. The *stx*-encoding prophage integration site (SBI) of *yehV* was shared among these three genomes, belonging to sub-lineage Ia (Table 3.1, Figure 3.3 and Figure 3.4). Both strain Sakai and EDL933 are temporally and geographically distinct from UK strain E30228. EDL933 was the cause of an outbreak associated with contaminated hamburgers in the USA in 1982^[16]. Strain Sakai is self-titled due to the outbreak in Japan associated with contaminated radish sprouts in 1996^[181].

Genomes belonging to sub-lineage IIc that were publicly available were associated with foodborne outbreaks in the UK^[68,139]. The *stx1a*-encoding prophages from these three isolates cluster together based on genetic distance but were genetically distant from the prophages described above in sub-lineage Ia genomes.

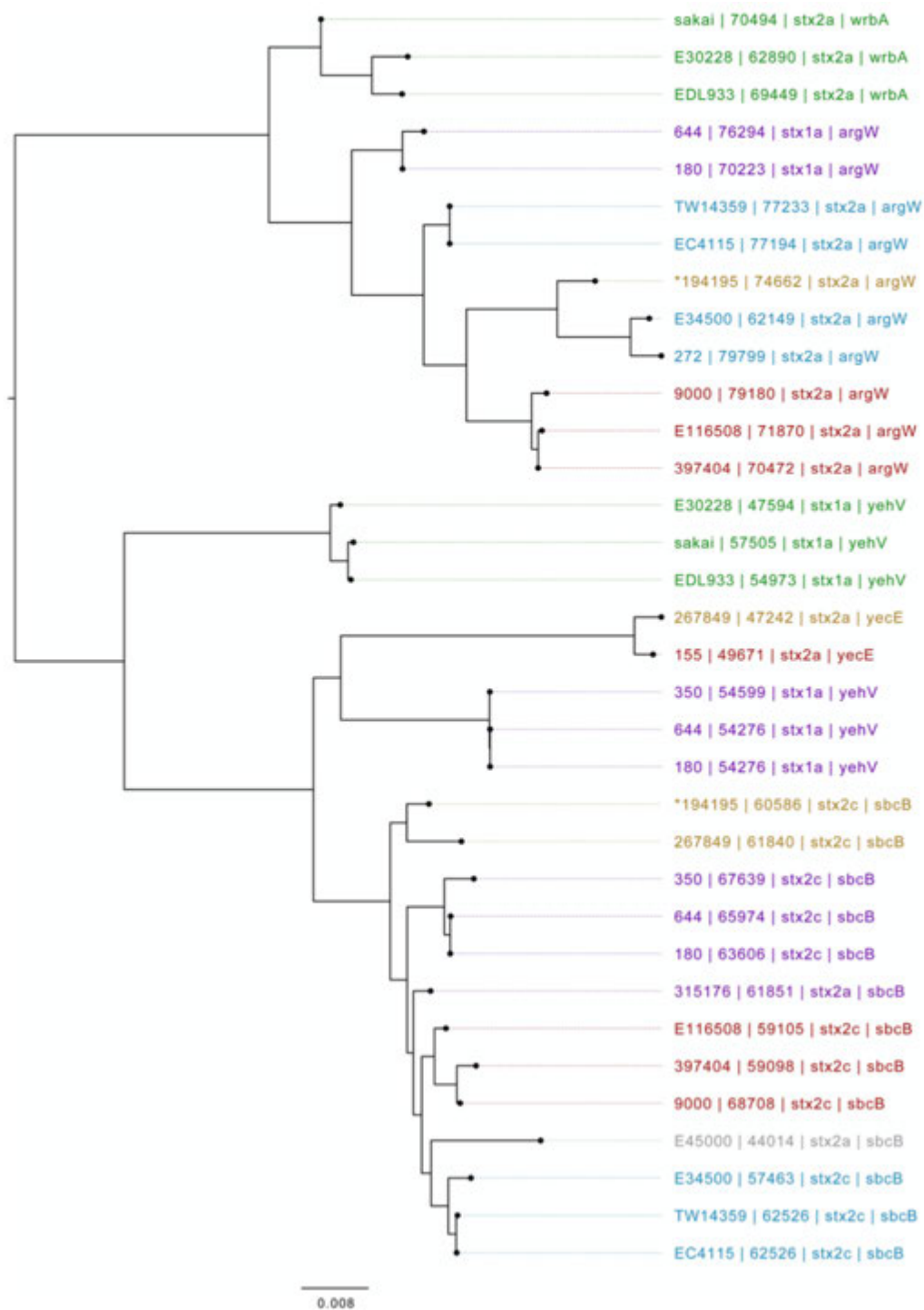


Figure 3.3 – Mid-rooted neighbour-joining tree of *stx*-encoding prophages based on Jaccard distance produced from Mash. Strains are annotated with strain ID, length (bp), *stx* subtype and SBI. Strains are coloured by sub-lineage: green, Ia; red, Ic; blue, I/IIa; grey, I/IIb; orange, IIa; black, IIb; purple, IIc. Scale bar indicates Jaccard distance. * denotes the non-domestic strain sequenced in this project 194195.

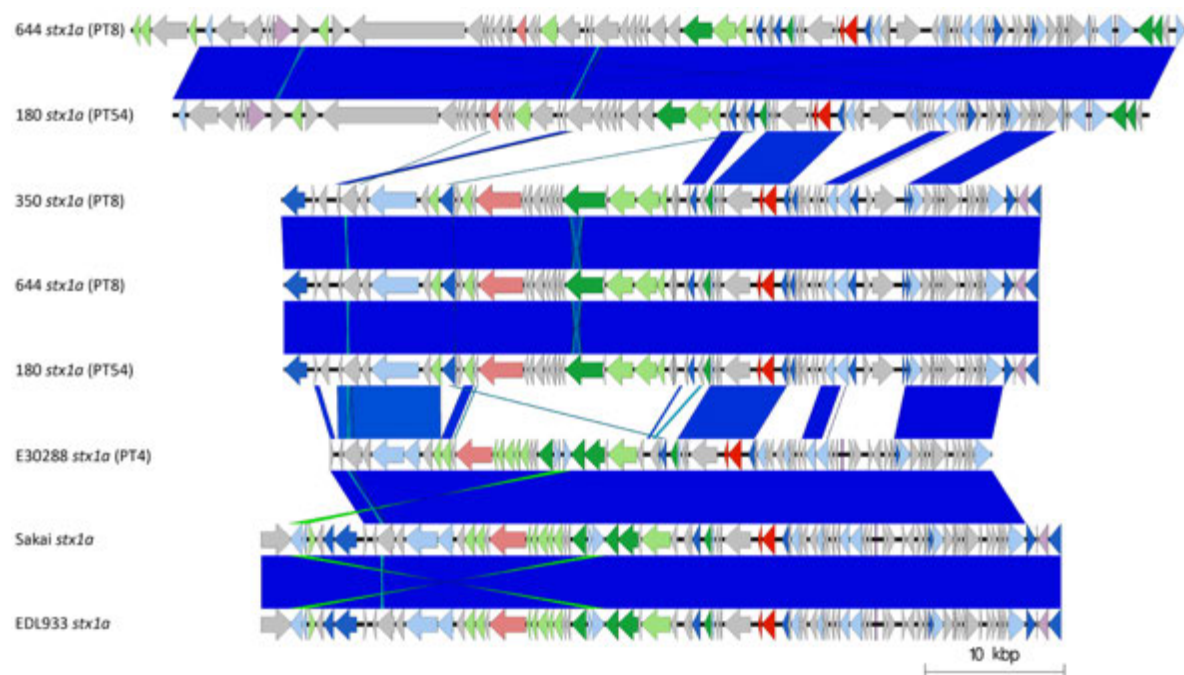


Figure 3.4 – Easyfig plot showing an alignment comparing the *stx1a*-encoding prophages from strains 644 (n=2 *stx1a*-encoding prophages), 180 (n=2 *stx1a*-encoding prophages), 350, E30288, EDL933 and Sakai. Arrows indicate gene directions. *stx* genes are shown in red; recombination/replication genes are shown in light blue; regulation-associated genes are shown in dark blue; effector genes are shown in pink; phage structure- and lysis-associated genes are shown in light and dark green, respectively; tRNAs are shown as purple lines; finally, hypothetical genes are shown in grey.

3.3.3 Comparison of the *stx2c*-encoding prophages.

Prophages harbouring the *stx2c* gene were observed in ten genomes from four different sub-lineages (Ic, I/IIa, IIa and IIc) (Table 3.1). Each of these *stx2c*-encoding prophages cluster together based on genetic distance and were all part of a single clade (Figure 3.3). The sequences aligned across the length of the prophages except for a few small structural variations (Figure 3.5). All *stx2c*-encoding

prophages shared the same SBI, designated *sbcB*. The *stx2c*-encoding prophages from strains within the same sub-lineage were more similar based on genetic distance than *stx2c*-encoding prophages in strains from different lineages (Table 3.1, Figure 3.3 and Figure 3.5). These strains were isolated over a wide time frame from 1983 to 2016, and a wide geography including the UK, Ireland and the USA, providing further evidence that *stx2c*-encoding prophages show a high level of similarity across lineages, time and geographical regions (Table 1) (Shaaban et al 2016).

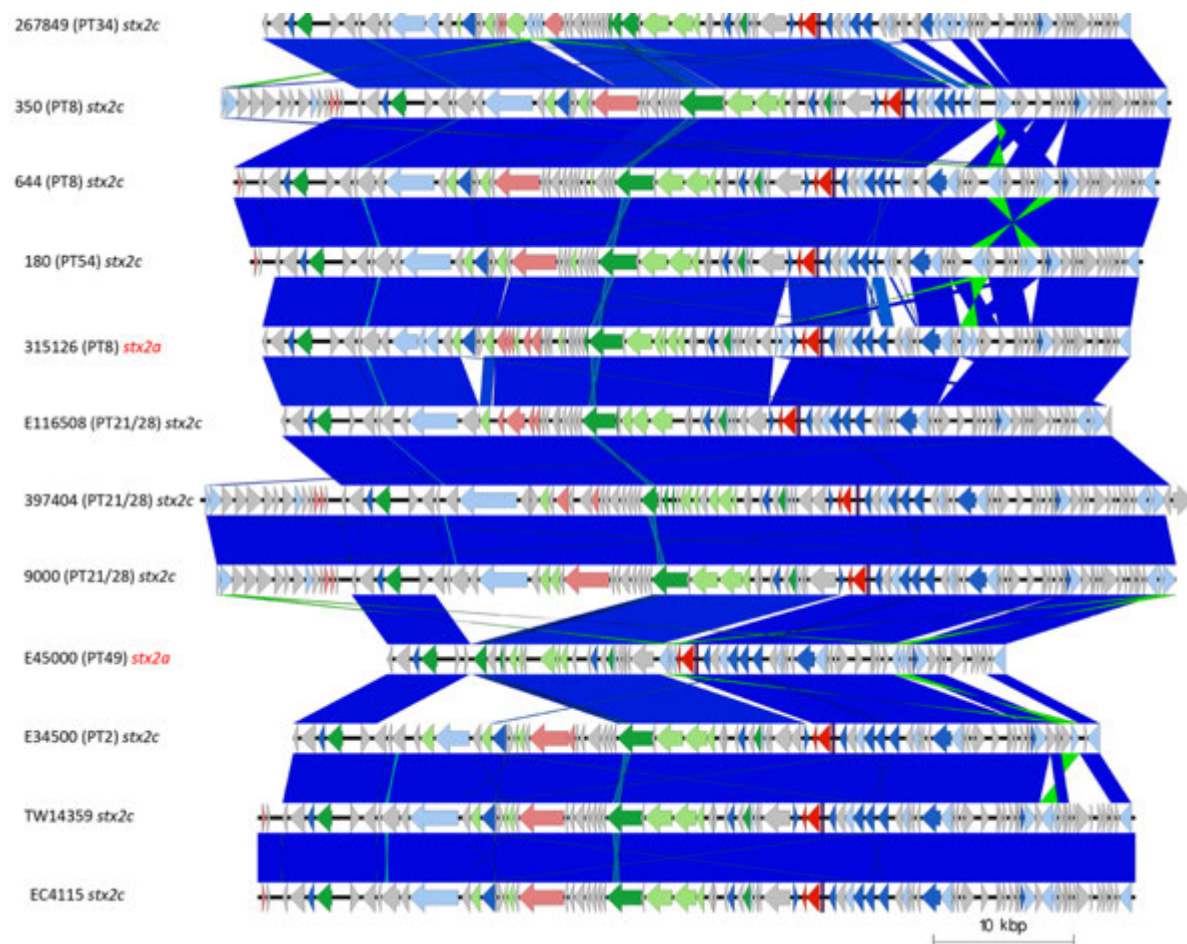


Figure 3.5 – Easyfig plot showing an alignment comparing the *stx2c*-encoding prophages from all samples in the study, including two *stx2a*-encoding prophages

that are associated with a *stx2c*-encoding prophage structure named in red (315126 and E45000). Arrows indicate gene directions. *stx* genes are shown in red; recombination/replication genes are shown in light blue; regulation-associated genes are shown in dark blue; effector genes are shown in pink; structure- and lysis-associated genes are shown in light and dark green, respectively; tRNAs are shown as purple lines; finally, hypothetical genes are shown in grey.

3.3.4 Comparison of the *stx2a*-encoding prophages.

Of samples sequenced in this project or publicly available genomes, *stx2a*-encoding prophages were the most frequently detected. Strains that shared sub-lineage, PT and geography harboured more similar *stx2a*-encoding prophages based on genetic distance. Three examples included (a) two sub-lineage I/IIa genomes from the USA, (b) two UK PT21/28 sub-lineage Ic genomes, (c) finally two UK PT2 sub-lineage I/IIa genomes (Table 3.1, Figure 3.3 and Figure 3.6).

Temporally related but geographically distant, genomes E30228 and EDL933, both sub-lineage Ia, encoded similar *stx2a*-encoding prophage (Table 3.1 and Figure 3.3). As did isolates 155 (sub-lineage Ic PT32) and 267849 (sub-lineage IIa PT34), which were unrelated temporally and geographically. When compared to other *stx*-encoding prophages (*stx1a*- and *stx2c*-encoding), the *stx2a*-encoding prophages found in 12 of the isolates in this project showed greater genetic diversity.

Comparison of the *stx2a*-encoding prophages in sub-lineage I/IIb showed genetic similarity to the prophages commonly found to encode *stx2c* genes (Figure 3.3 and Figure 3.5). *stx2a*-encoding prophages associated with sub-lineage I/IIb shared the same SBI as sub-lineage I/IIa *stx2c*-encoding prophages, specifically the *sbcB* gene.

The *stx2c*-encoding prophage in lineage Ic had high sequence similarity to *stx2c*-encoding prophages in the other genomes analysed in this study and shared the same SBI, *sbcB* (Table 3.1, Figure 3.2 and Figure 3.5). However, the *stx2a*-encoding prophage acquired by sub-lineage Ic differed genetically from those found in the three sub-lineages circulating in the previous decade (Ia, I/IIa and I/IIb) (Tables 3.1, Figure 3.2 and Figure 3.6).

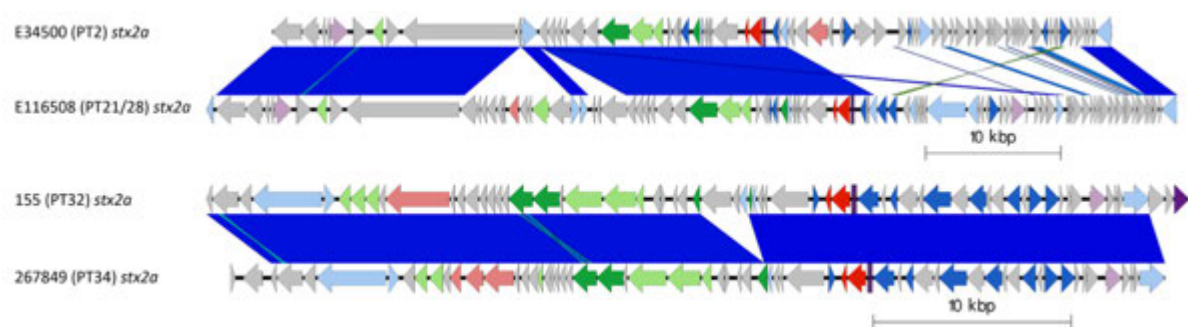


Figure 3.6 – Two Easyfig plots showing alignments comparing the *stx2a*-encoding prophages from E45000 with E116508 (above) and 155 and 267849 (below), in descending order. Arrows indicate gene directions. *stx* genes are shown in red; recombination/replication genes shown in light blue; regulation-associated genes are shown in dark blue; effector genes are shown in pink; structure- and lysis-associated genes are shown in light and dark green, respectively; tRNAs are shown as purple lines; finally, hypothetical genes are shown in grey.

3.3.5 Comparison of the prophages of the non-domestic strains.

Importation of STEC O157:H7 strains from outside the UK via contaminated food products is a threat to public health. In 2016, a large national outbreak of STEC O157:H7 PT34 harbouring both *stx2a* and *stx2c* belonging to sub-lineage IIa

occurred in the UK ^[107]. Contaminated red Batavia salad leaves from a non-domestic source was the most plausible vehicle of infection determined by epidemiological investigations. The *stx2a*-encoding prophage was noted to be different from all the *stx2a*-encoding prophages identified in the five major UK sub-lineages (Figure 3.3). This *stx2a*-encoding prophage however did share sequence similarity with another *stx2a*-encoding prophage from a PT32 STEC O157:H7 from sub-lineage Ic. This sample was epidemiologically linked to cases of severe gastrointestinal disease in Ireland ^[80] (Figure 3.3 and Figure 3.6). Unlike the previously described *stx2a*-encoding prophage which almost all shared the same SBI of *argW*, the *stx2a*-encoding prophage in both strains share the SBI *yecE*. This prophage also had similarity to the *stx2a*-encoding prophage found in a strain of STEC O55:H7 causing recurrent, seasonal outbreaks of HUS in England ^[189].

The Nanopore sequenced non-domestic strain in this project showed that it harboured a total of 19 prophages of which two were *stx2a*- and *stx2c*-encoding prophages (Table 3.1, Figure 3.3). Analysis of the *stx2c*-encoding prophage determined the SBI was also *sbcB*, and this prophage clustered on the same branch as *stx2c*-encoding prophages from strains from different time frames, geographical regions, and sub-lineages. However, it was most closely related to *stx2c*-encoding prophages from sub-lineages IIa (Figure 3.3).

The *stx2a*-encoding prophage associated with the non-domestic strain sequenced in this project (strain 267849), shared the same SBI as most observed *stx2a*-encoding prophages, *argW*. This prophage was most closely related to a *stx2a*-encoding prophage acquired by a strain belonging to lineage I/II that caused the first outbreak of STEC-HUS in the West Midlands in England in the early 1980s (Figure 3.3) ^[35,170].

3.4 Chapter 3 discussion.

3.4.1 *stx1a*-encoding prophages.

Previous analysis of isolates of sub-lineage Ia harbouring *stx1a*-encoding prophage indicate the *stx1a*-encoding prophage is likely ancestral and inherited vertically ^[80]. Strains analysed in this study corroborate this finding by encoding similar *stx1a*-encoding prophages, despite being isolated at different times and geographical locations.

As previously described ^[81,189], the *stx1a*-encoding prophages from samples from sub-lineage IIc, two of these strains (strain 180 PT54 and strain 664 PT8), linked to an outbreak in Northern Ireland in 2013 ^[139], had an additional but different *stx1a*-encoding prophage within each chromosome (Figure 3.3 and 3.4). As short-read sequencing cannot discriminate prophage content by read-based methods, the second *stx1a*-encoding prophage was not discovered until Nanopore sequencing was performed. Thus, in total, three different *stx1a*-encoding prophages, in two different lineages (Ia and IIc), were identified (Figure 3.3 and 3.4).

3.4.2 *stx2c*-encoding prophages.

The fact that *stx2c*-encoding bacteriophage show little variation is consistent with the hypothesis that the *stx2c*-encoding bacteriophage was acquired prior to the global dissemination and regional expansions of STEC O157:H7 ^[80] and has subsequently been maintained by vertical inheritance in the population across time and space.

3.4.3 *stx2a*-encoding prophages.

Strains of STEC O157:H7 causing HUS in the UK during the 1980s mainly belonged to two UK sub-lineages (Ia and I/IIa) and strains from both sub-lineages harboured *stx2a* ^[18,35] (Figure 3.1). However, during the 1980s, the number of primarily PT4 and PT1 sub-lineage Ia strains declined and a new sub-lineage I/IIb PT49, emerged (Figure 3.1). The emerging sub-lineage I/IIb PT49 strain also harboured a *stx2a*-encoding bacteriophage but one that was structurally different from either of the two *stx2a*-encoding prophages found in the representative isolates from the early contemporary sub-lineages Ia and I/IIa (Figure 3.3).

All three of the dominant 1980s sub-lineages (Ia, I/IIa and I/IIb) declined during the 1990s and a new sub-lineage emerged. STEC O157:H7 PT32 harbouring *stx2c*, belonging to sub-lineage Ic had been circulating in UK and Irish cattle populations for many decades but had not been linked to cases of human disease ^[80]. However, there was a change in the phage type from PT32 to PT21/28 following acquisition of a *stx2a*-encoding prophage (SBI is *argW*) ^[80,81], which was different in structure to those described above. The closest match to this *stx2a*-encoding prophage was one from sub-lineage I/IIa PT2 but showed differences in the genes associated with bacteriophage replication and recombination (Figure 3.6). Subsequently, sub-lineage Ic became the most common STEC O157:H7 sub-lineage causing gastrointestinal disease leading to HUS in humans in the UK for the next two decades.

More recently in the UK, the incidence of sub-lineage Ic PT21/28 has decreased but a new sub-lineage IIb PT8 has emerged again after acquiring a *stx2a*-encoding prophage ^[100]. Strains belonging to this sub-lineage have been linked to food-borne outbreaks more specifically linked to contaminated mixed-leaf salad ^[190], lamb-based meat products including sausages and mince, and an environmental exposure linked

to participation in a mud-based obstacle event ^[191]. Like the *stx2a*-encoding prophage described in sub-lineage I/IIb, the *stx2a*-encoding prophage in sub-lineage IIb exhibited high similarity to the *stx2c*-encoding prophages from sub-lineage IIb (Figure 3.3 and 3.5), with only minor alterations to the bacteriophage replication genes. This most likely occurred as a result of homologous recombination of the *stx2a* gene into a prophage previously encoding *stx2c*, with prophage structural genes acting as the region for homologous recombination to occur. This is also corroborated by the *stx2a*-encoding prophage in sub-lineage IIb integrating at *sbcB* associated with *stx2c*-encoding prophages (Table 3.1, Figure 3.3 and Figure 3.5). When compared to the *stx2c*- and *stx1a*-encoding prophages, the *stx2a*-encoding prophage exhibits greater genetic diversity (Figure 3.3, Figure 3.4, Figure 3.5 and Figure 3.6). In certain countries, the regional expansion of specific sub-lineages, or clades within sub-lineages, has involved acquisition of a *stx2a*-encoding prophage ^[80,91,170]. However, the limitation that public health agencies are more likely to receive strains that harbour *stx2a* due to its association with more severe disease and increased likelihood that these cases will present to healthcare.

3.5 Chapter 3 conclusions.

By utilising long-read Nanopore sequencing to generate complete genomes of STEC O157:H7, it has allowed the ability to characterise prophages harboured by the chromosome of its bacterial host. Across all *stx*-encoding prophages, *stx2a*-encoding prophages exhibited a higher level of diversity and there was little evidence of temporal or geographical patterns of relatedness, or of intra-UK transmission of *stx2a*-encoding prophage between indigenous strains. The *stx2a*-encoding

prophages in the UK lineages associated with severe disease, appear to be acquired independently and most likely from different geographical and/or environmental sources.

Finally, compared to *stx2a*-encoding prophages, *stx2c*-encoding prophages exhibit a high level of similarity across lineages, geographical regions and time, and have likely been maintained and inherited vertically.

These data provide supporting evidence for the existence of a dynamic environmental reservoir of *stx2a*-encoding prophages which may circulate under the public health surveillance radar. This poses a threat to public health due to their potential for integration into competent, indigenous UK sub-lineages (or vice-versa) of STEC O157:H7.

Monitoring the transmission of strains of STEC O157:H7 on a global scale and tracking the exchange of *stx2a*-encoding phage between imported and indigenous strains, may provide an early warning of emerging threats to public health.

Chapter 4:

Revealing the intra-outbreak accessory
genome variation of Shiga toxin-producing
E. coli (STEC) O157:H7.

4.1 Chapter 4 background.

In 2015, Public Health England (PHE), now UK Health Security Agency (UKHSA), implemented high-throughput sequencing for the public health surveillance of gastrointestinal pathogens, which included STEC O157:H7. The final analysis of the data produced involved determining strain relatedness of STEC O157:H7 strains by the detection of SNPs. This is accomplished by aligning the Illumina reads to a single reference genome per each clonal complex. These results are used to identify linked cases and outbreaks of infectious disease ^[94,98]. As this process is read-based with no assembly involved, this means all repetitive and homologous features, including prophages, are lost from the analysis, and little is known about the variation in prophage content of STEC O157:H7 genomes.

STEC O157:H7 has a larger accessory genome than more common strains of *E. coli*, this is mainly due to approximately 10–15% (approximately 550,000-825,000 bp) of the genome comprised of prophage and prophage-like regions ^[111,127].

Therefore, analysis of prophage content, loss and acquisition of bacteriophage and chromosomal structural rearrangements within prophage regions contributes to our understanding of the evolutionary history, virulence and potentially the likely source and transmission of this zoonotic, foodborne pathogen. Oxford Nanopore

Technologies, the whole genome sequencing platform, has been shown to achieve complete, closed contig *de novo* assemblies and this can facilitate more complete characterisation of the accessory genome including prophage regions ^[192,193,194].

This chapter aims to further our knowledge on the dynamic nature of STEC O157:H7 genome within an outbreak setting. Four UK outbreaks underwent Nanopore sequencing including those associated with foodborne transmission, animal contact,

eating and drinking of raw or contaminated food or milk and environmental transmission.

- An outbreak associated with contaminated raw pet food (n=4 strains).
- An outbreak associated with consumption of raw drinking milk (n=23 strains) [59].
- An outbreak associated with contaminated mud-based obstacle course (n=12 strains) [191].
- An outbreak associated with animal contact during a lambing event (n=10 strains) [61].

4.2 Chapter 4 methods.

4.2.1 Strain selection and preparation.

A breakdown of the selected outbreaks including total strains sequenced for each outbreak are as follows:

- An outbreak associated with contaminated raw pet food (n=4 strains).
- An outbreak associated with consumption of raw drinking milk (n=23 strains).
- An outbreak associated with contaminated mud-based obstacle course (n=12 strains).
- An outbreak associated with animal contact during a lambing event (n=10 strains).

All strains selected were cultured from long-term storage at UK Health Security Agency (UKHSA) [known previously as Public Health England (PHE)] using 4ml of double-strength nutrient broth (Oxoid) and incubated at 37°C overnight.

4.2.2 Illumina-based DNA extraction, library preparation, sequencing and data processing.

Genomic DNA was extracted from cultures of STEC O157:H7 using the QIAAsymphony system (Qiagen, Hilden, Germany). The sequencing library was prepared using the Nextera XP kit (Illumina, San Diego, USA) for sequencing on either the HiSeq 2500 and NextSeq 1000 instruments (Illumina, San Diego, USA), each run with the fast protocol. FASTQ reads were processed using Trimmomatic v0.27^[140] as previously described^[195] to remove sequencing barcode/index adapters. Also, bases from reads were removed from the trailing end of that fell

below a PHRED score of 30. If the read length post-trimming was less than 50, the read and its pair were discarded.

4.2.3 Nanopore-based DNA extraction, library preparation, sequencing and data processing.

High-molecular weight (HMW) genomic DNA was extracted and purified using the Revolugen Fire Monkey HMW DNA extraction kit (RevoluGen, UK) to manufacturer's instructions and Qiagen genomic tips, midi 100/G (Qiagen), with minor alterations including no vigorous mixing steps (mixing performed by inversion instead) and elution into 100µl double processed nuclease-free water (Sigma-Aldrich). Each gDNA extract was quantified using a Qubit and the HS (high sensitivity) dsDNA assay kit (ThermoFisher Scientific, Waltham, USA) to manufacturer's instructions. Library preparation was performed using the rapid barcoding kit (SQK-RBK004) with each extract being barcoded/indexed via transposase-based tagmentation. Sequencing was performed on the MinION Mk1B for 48 hours as previously described ^[170,196].

Raw FAST5 data was produced from the sequencer before basecalling and de-multiplexed using Guppy (versions v3.2.4 through to v3.2.10 all using the FAST model) (Oxford Nanopore Technologies, Oxford, UK) into FASTQ format. De-multiplexing was re-performed using Deepbiner v0.2.0 ^[182], sequencing run and QC metrics were generated using Nanoplot v1.8.1 ^[141]. Read trimming was performed using Porechop v0.2.4 ^[142] to remove the Nanopore y-adaptor and each barcode adaptor. Finally, read filtering using Filtlong v0.2.0 ^[143] with the following parameters, min_length=1000 bp, keep_percent=90 and target_bases=275 Mbp, to generate

approximately 50x coverage of the STEC genome with the longest and highest-quality reads.

4.2.4 *De novo* assembly, correction, reorientation, and annotation.

The filtered Nanopore FASTQ file with the 50x coverage of longest reads were assembled using Flye v2.8 and ^[144] with the minimum overlap length (-m) set to 10000 and the --meta component enabled. Correction (polishing) of the assemblies was performed in a three-step process. Firstly, using Nanopolish v0.11.3 ^[145] with the parameters set to --methylation-aware=dcn,dam, --min-candidate-depth=10 and --min-candidate-frequency=0.1 using Nanopore reads and a self-alignment generated from Minimap2 v2.7 ^[146] and Samtools v1.7 ^[147]. Secondly, using Pilon v1.22 ^[148] with parameters set to --minmq=0, --minqual=0 and --mindepth=0.05 with each samples' equivalent Illumina reads. Finally Racon v1.3.3 ^[150] with parameters set to --error-threshold=0.3 and --quality-threshold=10 again using each samples' equivalent Illumina reads. Nanopore based alignments were generated using Minimap2 v2.17 ^[146] and Samtools v1.7 ^[147]. Illumina based alignments using BWA-MEM v0.7.17 ^[149] and Samtools v1.7 ^[147].

If each assembly contig was circularised and closed, they were re-orientated to start at the *dnaA* gene (GenBank accession no. NC_000913) from *E. coli* K-12, using the --fixstart parameter in Circlator v1.5.5 ^[184]. Prokka v1.13 ^[155] was used to annotate the final assemblies.

4.2.5 Prophage detection, excision, comparison and generation of neighbour-joining trees.

Prophage sequences were detected and extracted from each samples' chromosome as described in Shaaban *et al* 2016 ^[81] and Yara *et al* 2020 ^[170]. Firstly, Propi v0.0.1 ^[81] was used to detect and extract prophages from each sample's corrected assembly. Assemblies were then manually curated for prophage detection and excision followed by prophage annotated using Prokka v1.13 ^[155]. Mash v2.2 ^[186] was used to sketch (sketch length 1000, kmer length, 21) all extracted prophages in the samples sequenced for each outbreak and in total in this project. The pairwise Jaccard distance between the prophages was calculated and several neighbour joining trees generated. Firstly, for *stx*-encoding prophages only, secondly for all prophages within a given outbreak and finally all prophages vs those from strain Sakai to act as a well described reference prophage set. Trees were visualised and annotated using FigTree v1.4.4 ^[187]. Prophages and chromosomes were also aligned and compared using Easyfig v2.2.5 ^[188].

4.2.6 Variant calling and masking, SNP typing and generation of maximum-likelihood phylogenetic trees.

For reference-based variant calling both Illumina and Nanopore FASTQ reads were mapped/aligned to the Sakai STEC O157 reference genome (BA000007) using BWA v0.7.3 ^[149] and Minimap2 v2.17 ^[146] respectively with the use of Samtools v0.7.17 ^[147]. Variant calling was produced using GATK v2.6.5 UnifiedGenotyper ^[163]. Variants that had a high-quality SNP ($\geq 90\%$ for Illumina) $\geq 80\%$ for Nanopore] % per-base consensus, minimum depth of 10 reads and mapping quality (MQ) ≥ 30 in at least one isolate were extracted for further analysis. Any variants called at positions that were within the known prophages in Sakai were masked from further analyses. 5-

methylcytosine positions were identified using Nanopolish V0.11.3 ^[145] and then masked from the Nanopore VCFs as described in Greig *et al* 2019 ^[171] and Chapter 2 of this thesis.

The maximum likelihood phylogenetic trees were constructed by RAxML v8.1.17 ^[168] using an alignment generated from SnapperDB ^[94] that recombination had been accounting for by Gubbins v2.00 ^[167]. Visualisation of the phylogenetic tree was performed using FigTree v1.4.4 ^[187].

To detect false positive/negative SNPs called by Illumina reads, discrepant variant positions between Illumina and Nanopore relative to the reference genome were extracted. Those variants that were called in paralogous sequences that also had a lower-than-average mapping quality were then masked in the alignment. To be included in the masking process the false called variant must be called the same in all the alignments of at least 50% of samples in the study.

4.2.7 Detection and characterisation of chromosomal structural variation.

Chromosome synteny was compared by aligning outbreak sample chromosomes using Easyfig v2.2.5 ^[188]. Once samples in one chromosome were aligned, structural variation could be determined and further characterised using Artemis v18.1.0 ^[197]. To determine if there were multiple isoforms within each sample's Nanopore reads (FASTQ), the FASTQ for an outbreak sample where the chromosome showed one isoform was aligned to a finalised assembly of a sample with a different chromosomal isoform using Minimap2 v2.17 ^[146] and Samtools v0.7.17 ^[147]. Using Tablet v1.17.08.17 ^[166], the alignments were visualised and the breakpoints at each

isoform where identified. Once breakpoints were identified relative to each isoform, those positions were used with Samtools v0.7.17 ^[147] view to isolate reads that must align across both ends of each prophage breakpoint. Any reads that did align across a given set of breakpoints must share the same size as it exists in the FASTQ file and not clipped, to be considered.

4.2.8 *In silico* plasmid typing and characterisation.

The plasmid replicon for each non-chromosomal contig within each sample's final assembly was compared to PlasmidFinder's v2.1 Enterobacteriaceae reference database ^[198]. An alignment was generated and visualised using Easyfig v2.2.5 ^[188] relying on BLASTn v2.9 ^[199]. BLASTn parameters used were minimum_identity = 90% and minimum_length_hit = 100bp.

4.3 Chapter 4 results.

4.3.1 Outbreak associated with contaminated raw pet food.

4.3.1.1 Assembly metrics, variant calling and phylogeny.

The finalised assemblies for each of the four isolates sequenced in this outbreak resolved into two contigs, the larger being the chromosome and the other was the IncFIB pO157 plasmid. The chromosome size for cases A to D were 5,486,665bp, 5,487,004bp, 5,486,935bp and 5,424,337bp respectively. The size of the second contigs (replicon type) in each sample were 91,449bp (IncFIB), 91,445bp (IncFIB) 91,445bp (IncFIB) 91,443bp (IncFIB) and for cases A to D respectively. Sample B contained an extra 57,390bp IncI2 plasmid which other samples did not.

In total, variant calling using GATK and SnapperDB identified between zero and six SNPs when comparing each sample with both sequencing technologies to one another (Table 4.1).

Sample	Case A ONT	Case B ONT	Case C ONT	Case D ONT	Case A Illumina	Case B Illumina	Case C Illumina	Case D Illumina
Case A ONT	-	0	0	0	5	6	5	5
Case B ONT	0	-	0	0	5	6	5	5
Case C ONT	0	0	-	0	5	6	5	5
Case D ONT	0	0	0	-	5	6	5	5
Case A Illumina	5	5	5	5	-	1	0	0
Case B Illumina	6	6	6	6	1	-	0	0
Case C Illumina	5	5	5	5	0	0	-	0
Case D Illumina	5	5	5	5	0	0	0	-

Table 4.1 – Table showing the pairwise genetic distance matrix (SNPs) for both Illumina and Nanopore technologies for all four cases.

Concerning the Illumina generated sequence data, there was only a single SNP detected between each sample. This individual SNP was from sample B at position 2,578,517 relative to the reference genome within a gene that encodes a proton conductor (Table 4.1, Figure 4.1). For the Nanopore-based variant calling, there were no SNPs detected between all four samples relative to the reference genome (Table 4.2).

In total there were five discrepant SNPs between Nanopore and Illumina variant calls, relative to the reference genome. One of the five discrepant variants, one was from the Illumina dataset, the remaining four were from the Nanopore dataset (Table 4.2).

Short-read Illumina sequences demonstrated ambiguous alignment to paralogous sequences in the reference genome which problematically led to the introduction of a single false positive variant calls. The remaining four false positive or false negative SNPs were generated from a known systemic error during the base-calling process of homopolymer regions in Nanopore sequencing.

Position in reference genome	Base in reference genome	Base in Illumina data	Base in Nanopore data
270,595	C	A	C
379,516	A	G	A
2,033,176	T	G	T
4,709,195	A	A	G
4,901,209	A	A	G

Table 4.2 – Table showing the position of the variants between the two sequencing technologies for all four samples.

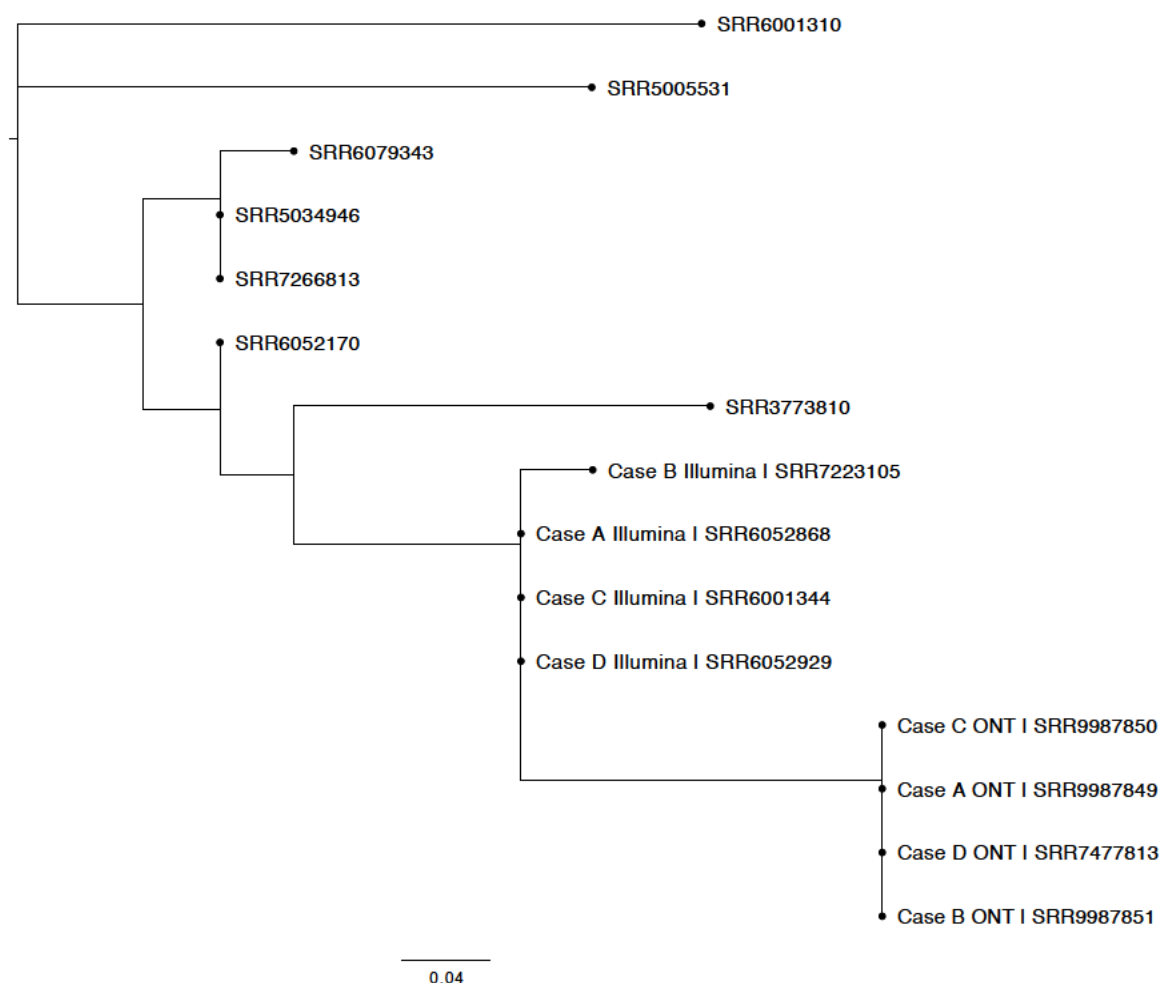


Figure 4.1 – Maximum-likelihood phylogenetic tree of the four outbreak samples within this outbreak and the seven most closely related isolates in the UKHSA archive to show context. Both data from Illumina and Nanopore sequencing technologies for the four outbreak cases (A–D) is also show.

4.3.1.2 Analysis of prophage content of the outbreak isolates and comparison.

Each sample produced complete, closed chromosomes thus facilitating the interrogation of the chromosome for prophage sequences. In total there were 16 prophage regions, of which 15 were shared between the four outbreak isolates with

sample D containing an extra prophage. The prophage regions ranged in size from 8kbp to 145kbp (Table 4.3 and Figure 4.2).

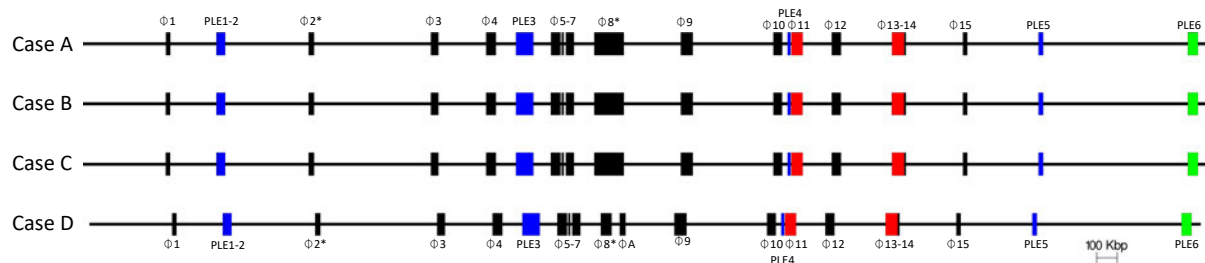


Figure 4.2 – Diagram showing linearised chromosomes and the loci of the 15 prophages (16 in sample D). Red prophages indicate *stx*-encoding prophages. Blue indicates prophage like elements. Green indicates the locus of enterocyte effacement (LEE).

Specifically, seven of the total 16 prophages showed almost identical similarity to the prophages obtained from the Sakai strain, ranging from 98–100% nucleotide identity and coverage (Table 4.3 and Figure 4.3). Prophages 3, 4, 5, 7, 10, 14 and 15 from the four outbreak strains matched with the Sakai strain prophages Sp3, Sp4, Sp6, Sp8, Sp14, Sp16 and Sp17 respectively and shared the same bacteriophage integration site for each prophage (*ybhC*, *yccA*, *potC*, *icd*, *serU*, *argW* and *ssrA*, again respectively) (Table 4.3 and Figure 4.3).

Prophage detected	Gene 5' to prophage	Gene 3' to prophage	Size (bp) in sample A	Size (bp) in sample B	Size (bp) in sample C	Size (bp) in sample D	Position in sample A	Position in sample B	Position in sample C	Position in sample D
1	<i>lexA</i>	<i>aphA</i>	22,035	22,035	22,034	22,036	403,708-425,743	403,729-425,764	4,03,712-4,25,746	403,708-425,744
2 ^a	<i>tRNA-Thr (cgt)</i> ^b	<i>prgR</i>	26,826	26,823	26,823	26,821	1,101,413-1,128,239	1,101,440-1,128,263	1,101,421-1,128,244	1,101,414-1,128,235
3	<i>ybhC</i> ^b	<i>ybhB</i>	38,095	38,095	38,096	38,095	1,697,277-1,735,372	1,697,420-1,735,515	1,697,397-1,735,493	1,697,388-1,735,483
4	<i>yccA</i> ^b	<i>tRNA-Ser (tga)</i>	48,257	48,277	48,265	48,263	1,967,370-2,015,627	1,967,513-2,015,790	1,967,491-2,015,756	1,967,481-2,015,744
5	<i>potC</i> ^b	<i>potB</i>	47,889	47,679	47,879	47,896	2,283,891-2,331,780	2,284,054-2,331,733	2,284,019-2,331,898	2,284,012-2,331,908
6	<i>raxA</i> ^b	<i>phoQ</i>	10,436	10,436	10,436	10,436	2,337,022-2,347,458	2,336,975-2,347,411	2,337,140-2,347,576	2,337,150-2,347,586
7	<i>icd</i> ^b	<i>caeB</i>	40,992	40,991	40,992	40,991	2,356,230-2,397,222	2,356,183-2,397,174	2,356,348-2,397,340	2,356,358-2,397,349
8 ^a	<i>ompW</i> ^b	<i>rspR</i>	145,151	144,981	145,040	53,636	2,495,566-2,640,717	2,495,518-2,640,499	2,495,685-2,640,725	2,495,693-2,549,329
A	<i>hipA</i>	<i>ydeP</i>	/	/	/	28,513	/	/	/	2,589,222-2,617,735
9	<i>trpA</i>	<i>rspA</i>	59,339	59,340	59,322	59,863	2,918,047-2,977,386	2,917,834-2,977,174	2,918,061-2,977,383	2,855,169-2,915,032
10	<i>yodB</i>	<i>tRNA-Ser (cga)</i> ^b	43,681	43,694	43,684	43,690	3,370,042-3,413,723	3,369,847-3,413,541	3,370,056-3,413,740	3,307,170-3,350,860
11 (<i>stx2c</i>)	<i>yeeA</i>	<i>tnpA</i> ^b	55,650	56,910	55,652	55,655	3,457,661-3,513,311	3,456,222-3,513,132	3,457,677-3,513,329	3,394,797-3,450,452
12	<i>yehV</i>	<i>yehV</i> ^b	45,720	45,725	45,732	45,719	3,654,766-3,700,486	3,654,584-3,700,309	3,654,782-3,700,514	3,591,904-3,637,623
13 (<i>stx2a</i>)	<i>yfdC</i>	<i>argW</i> ^b	60,239	60,237	60,239	60,242	3,948,563-4,008,802	3,948,386-4,008,623	3,948,591-4,008,830	3,885,700-3,945,942
14	<i>argW</i> ^b	<i>lacY</i>	8,233	8,233	8,233	8,233	4,009,234-4,017,467	4,009,055-4,017,288	4,009,262-4,017,495	3,946,374-3,954,607
15	<i>ssrA</i>	<i>alpA</i> ^b	22,107	22,103	22,098	22,106	4,294,770-4,316,877	4,294,584-4,316,687	4,294,801-4,316,899	4,231,915-4,254,021
Prophage-like element	Gene 5'	Gene 3'	Size (bp) in sample A	Size (bp) in sample B	Size (bp) in sample C	Size (bp) in sample D	Position in sample A	Position in sample B	Position in sample C	Position in sample D
PLE1	<i>tRNA-Leu</i>	<i>int</i>	9,530	9,530	9,530	9,530	650,711-660,241	650,730-660,260	650,601-660,131	650,715-660,245
PLE2	<i>int</i>	<i>nanS</i>	34,465	34,981	34,465	34,465	660,260-694,725	659,763-694,744	660,150-694,615	660,264-694,729
PLE3	<i>ycdU</i>	<i>tRNA-Ser</i>	85,906	85,906	85,906	85,911	2,112,819-2,198,725	2,113,064-2,198,970	2,112,913-2,198,819	2,113,028-2,198,939
PLE4	<i>cobU</i>	<i>yeeB</i>	15,044	15,045	15,044	15,046	3,439,339-3,454,383	3,439,883-3,454,928	3,439,719-3,454,763	3,377,102-3,392,148
PLE5	<i>tRNA-Phe</i>	<i>pitB</i>	23,334	23,333	23,333	23,332	4,664,210-4,687,544	4,664,345-4,687,678	4,664,588-4,687,921	4,601,991-4,625,323
PLE6	<i>selC</i>	<i>yicL</i>	50,390	50,389	50,390	50,390	5,392,598-5,442,988	5,392,972-5,443,361	5,393,215-5,443,605	5,330,617-5,381,007

Table 4.3 – Table showing the size, chromosomal position and integration sites of all 16 prophages (top) and all six prophage-like elements (bottom) within each of the outbreak sample chromosomes. a refers to prophages that appear to be compound prophages (i.e. two or more prophages that are sequential with no chromosomal sequence between them, with intact integrase genes). b refers to the direction of the prophage and where the integrase gene (*intA*) is located.

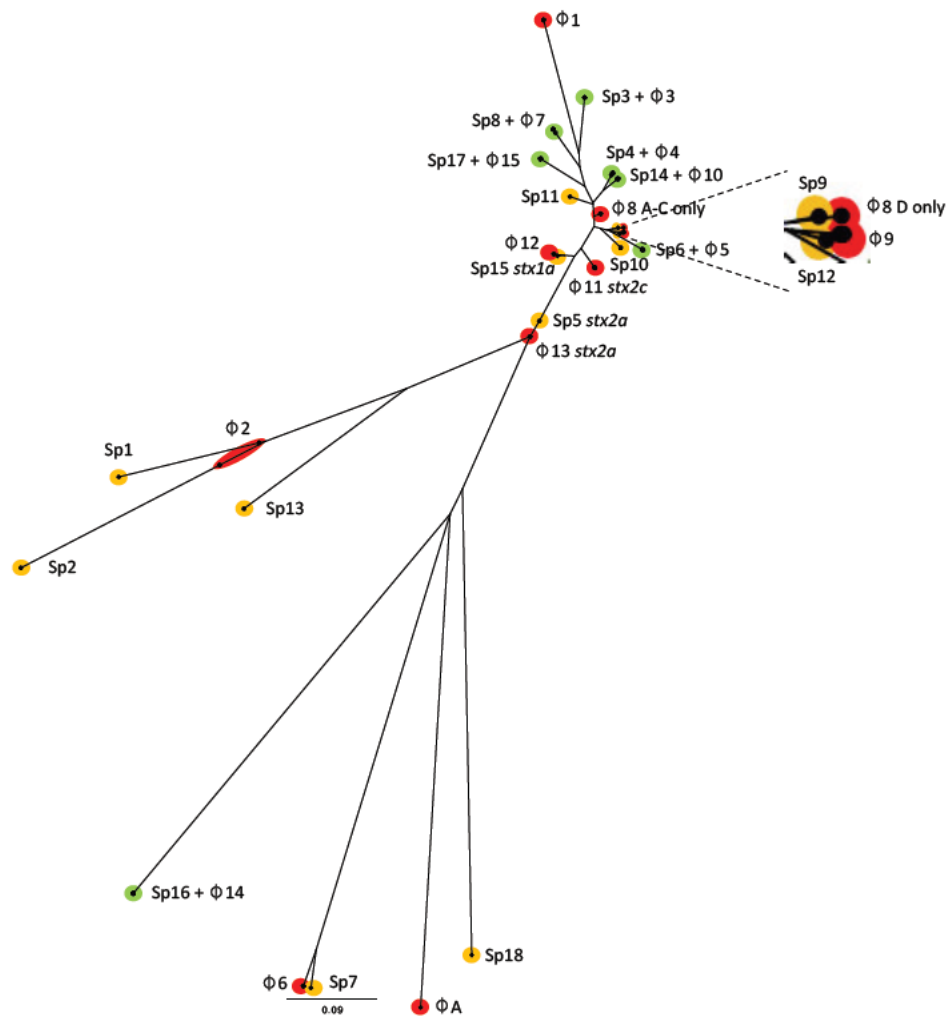


Figure 4.3 – Unrooted neighbour-joining tree based on Jaccard distances from mash showing genetic relationships for prophages from Cases A - D with prophages from BA000007 (Sakai). Green clusters show prophages shared in Cases A - D and Sakai, yellow clusters show prophages unique to Sakai only and red clusters show prophages unique to Cases A - D.

Case D had an extra prophage compared to the other three samples designated prophage “A”. This prophage was 28kbp in length and integrated at the *hipA* gene (Figure 4.2, Figure 4.3 and Table 4.3). There were nine prophages in the outbreak

isolates that were did not match or had <50% sequence similarity compared to the Sakai strain.

In the outbreak strain, *stx2a* and *stx2c* were encoded on prophages 11 and 13, inserted at *argW* and *tnpA*, respectively. For the *stx2c*-encoding prophage the known SBI site, *sbcB*, as previously described for a PT21/28 STEC O157:H7 [94,111,170,171] has been split by a short 2.7kbp insertion sequence (IS629) hence the SBI designation *tnpA* which encodes a transposase. The *stx2a*-encoding prophage detected in the outbreak strain had approximately 30% coverage but >97% nucleotide similarity with Sp5, which is the strain Sakai *stx2a*-encoding prophage. The regions of high similarity included the *stx*-encoding genes, *nin* region, Q-gene, DNA replication genes, recombination genes and the prophage structural regions differed including head, tail and tail fibres/tip genes.

4.3.1.3 Within-outbreak comparison of the prophage regions of the four outbreak isolates.

Once the chromosomes of the outbreak strains were aligned, two genome rearrangements/modification events were identified. Firstly, within prophage 2, where cases A and B differ from C and D. A 1,739 bp inversion was identified involving two prophage tail genes each surrounding a hypothetical gene (*yfdK*) (Figure 4.4).

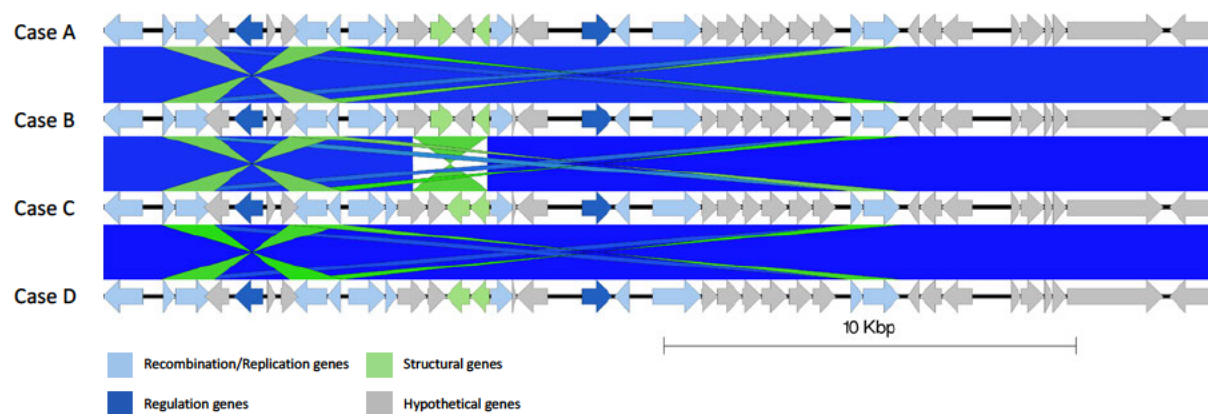


Figure 4.4 – Easyfig plot showing the alignment of prophage 2 in all cases order descending A–D, detailing a small inversion from samples A and B relative to samples C and D (light green inverted sequence), Blue denotes high sequence similarity. Arrow direction also indicates gene direction.

The second event was located in prophage 8 where a deletion event was observed in the sequence linked to case D with respect the other outbreak sequences (Figure 4.5). Prophage 8 is a large compound prophage (53kbp in the sequence linked to case D and 145kbp in the remaining samples) containing at least three separate prophages positioned sequentially without any chromosomal sequence separating them. The genome of the strain linked to case D has a 92kbp deletion (relative to the three other samples). The deletion contained almost two full prophage sequences, making up two sets of structural bacteriophage genes, regulatory genes, lysis genes and site-to-site recombination genes totalling 92kbp in length. Within the four outbreak strains, the prophage content was equivalent except for the deleted prophage sequence in prophage 8 and the acquisition of another prophage in case D (Figure 4.3 and Figure 4.5).

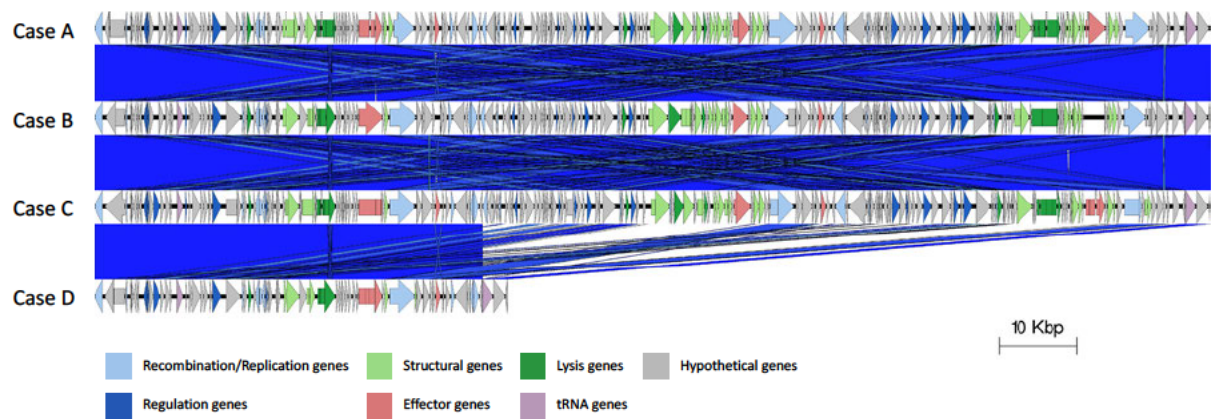


Figure 4.5 – Easyfig plot showing the alignment of prophage 8 in all cases order descending A–D. Showing a substitution in sample D. Arrow direction also indicates gene direction. Blue blocks denote regions of high sequence similarity.

4.3.2 Outbreak associated with consumption of raw drinking milk.

4.3.2.1 Genome features of long-read assemblies.

The assemblies of the sequences of the 23 isolates in this study all contained either two or three closed contigs, each supporting a single replicon (Table 4.4). The chromosome size of the isolates sequenced in this study (n=23) ranged from 5,507,151bp to 5,555,878bp in length, a maximum difference of 48,727bp and an average size of 5,553,177bp (Table 4.4).

Aligning and comparing the chromosomes of all 23 isolates within the same 5-SNP single linkage cluster led to the discovery of a large-scale recombination (LCR) event present in one of the strains isolated from cattle (432301) (Figure 4.6). The large-scale recombination event was characterised by a 650kbp inversion between prophages 5 (*potC*) and 7 (*yebW*) (Figure 4.6). At either edge of the inversion, prophages 5 (*potC*) and 7 (*yebW*) both have a 10.05kbp homologous sequence containing prophage structure encoding genes such as tail proteins, host specificity

proteins and several hypothetical genes. The 10.05kbp homologous sequences in both prophages share a 97.7% sequence similarity.

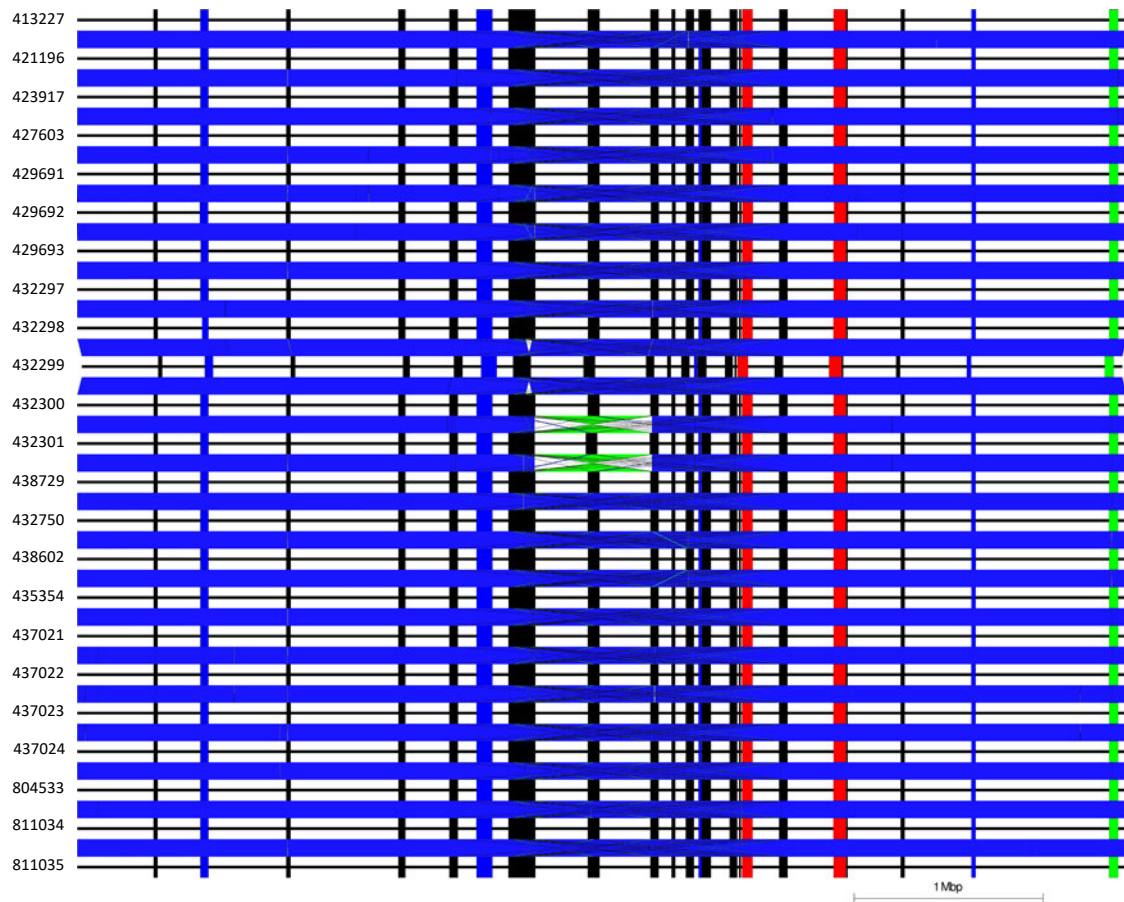


Figure 4.6 – Easyfig plot showing an alignment linearised chromosomes and loci of prophages in all samples sequenced in this study. *Stx*-encoding prophage, Red; Prophage-like region, Blue; Locus of Enterocyte Effacement (LEE), Green and other non-*stx*-encoding prophages, Black. Blue links denote regions of similarity, white links show regions of difference and green show regions of similarity with different directionality.

Strain ID	Source	# of contigs	Chromosome size (bp)	Total genome size (bp)	Plasmid 1 size (bp) and Inc group	Plasmid 2 size (bp) and Inc group	# of prophages
413227	Human	2	5,555,475	5,649,172	93,697 IncFIB	-	17
421196	Human	2	5,555,066	5,649,123	94,057 IncFIB	-	17
423917	Human	2	5,555,034	5,648,723	93,689 IncFIB	-	17
427603	Human ^a	2	5,555,537	5,649,596	94,059 IncFIB	-	17
429691	Milk	2	5,555,094	5,649,152	94,058 IncFIB	-	17
429692	Milk	2	5,558,855	5,652,943	94,088 IncFIB	-	17
429693	Milk	2	5,555,560	5,649,616	94,056 IncFIB	-	17
432297	Animal	2	5,554,755	5,648,807	94,052 IncFIB	-	17
432298	Animal	2	5,554,878	5,648,605	93,727 IncFIB	-	17
432299	Animal	2	5,507,151	5,601,209	94,058 IncFIB	-	17
432300	Animal	2	5,555,280	5,649,335	94,055 IncFIB	-	17
432301	Animal	2	5,554,492	5,648,544	94,052 IncFIB	-	17
438729	Animal	2	5,555,235	5,649,147	93,912 IncFIB	-	17
432750	Animal	2	5,555,568	5,649,625	94,057 IncFIB	-	17
438602	Animal	2	5,555,032	5,649,087	94,055 IncFIB	-	17
435354	Human	3	5,555,067	5,735,080	94,051 IncFIB	85,962 (IncI1-γ)	17
437021	Human ^a	2	5,555,314	5,649,370	94,056 IncFIB	-	17
437022	Human ^b	2	5,554,737	5,648,789	94,052 IncFIB	-	17
437023	Human ^b	2	5,554,791	5,648,842	94,051 IncFIB	-	17
437024	Human ^a	2	5,555,270	5,649,330	94,060 IncFIB	-	17
804533	Human	2	5,554,866	5,650,235	95,369 IncFIB	-	17
811034	Human ^c	2	5,555,205	5,650,573	95,368 IncFIB	-	17
811035	Human ^c	2	5,554,812	5,650,179	95,367 IncFIB	-	17

Table 4.4 – Table detailing the ID of each strain, source, finalised assembly metrics, plasmid replicon type and a count of prophages.

Strain ID	CC11 Sub-lineage	Phage Type	Stx-profile	<i>Stx1a</i> prophage SBI and size (bp)	<i>Stx2a</i> prophage SBI and size (bp)	<i>Stx2c</i> prophage SBI and size (bp)	Reference	Assembly Accession no.
413227	lc	PT21/28	<i>Stx2a/2a</i>	-	<i>sbcB</i> (56,183) <i>argW</i> (66,541)	-	This study	CP088060-1
421196	lc	PT21/28	<i>Stx2a/2a</i>	-	<i>sbcB</i> (56,429) <i>argW</i> (66,540)	-	This study	CP088058-9
423917	lc	PT21/28	<i>Stx2a/2a</i>	-	<i>sbcB</i> (56,427) <i>argW</i> (66,536)	-	This study	CP088056-7
427603	lc	PT21/28	<i>Stx2a/2a</i>	-	<i>sbcB</i> (56,435) <i>argW</i> (66,542)	-	This study	CP088071-2
429691	lc	PT21/28	<i>Stx2a/2a</i>	-	<i>sbcB</i> (56,420) <i>argW</i> (66,537)	-	This study	CP088054-5
429692	lc	PT21/28	<i>Stx2a/2a</i>	-	<i>sbcB</i> (56,426) <i>argW</i> (66,536)	-	This study	CP088052-3
429693	lc	PT21/28	<i>Stx2a/2a</i>	-	<i>sbcB</i> (56,422) <i>argW</i> (66,540)	-	This study	CP088050-1
432297	lc	PT21/28	<i>Stx2a/2a</i>	-	<i>sbcB</i> (56,150) <i>argW</i> (66,452)	-	This study	CP088048-9
432298	lc	PT21/28	<i>Stx2a/2a</i>	-	<i>sbcB</i> (56,439) <i>argW</i> (66,539)	-	This study	CP088046-7
432299	lc	PT21/28	<i>Stx2a/2a</i>	-	<i>sbcB</i> (56,427) <i>argW</i> (66,537)]	-	This study	CP088044-5
432300	lc	PT21/28	<i>Stx2a/2a</i>	-	<i>sbcB</i> (56,430) <i>argW</i> (66,540)	-	This study	CP088042-3
432301	lc	PT21/28	<i>Stx2a/2a</i>	-	<i>sbcB</i> (56,428) <i>argW</i> (66,539)	-	This study	CP088040-1
438729	lc	PT21/28	<i>Stx2a/2a</i>	-	<i>sbcB</i> (56,441) <i>argW</i> (66,536)	-	This study	CP088069-70
432750	lc	PT21/28	<i>Stx2a/2a</i>	-	<i>sbcB</i> (56,433) <i>argW</i> (66,538)	-	This study	CP088067-8
438602	lc	PT21/28	<i>Stx2a/2a</i>	-	<i>sbcB</i> (56,379) <i>argW</i> (66,537)	-	This study	CP088038-9
435354	lc	PT21/28	<i>Stx2a/2a</i>	-	<i>sbcB</i> (56,441) <i>argW</i> (66,539)	-	This study	CP088064-6
437021	lc	PT21/28	<i>Stx2a/2a</i>	-	<i>sbcB</i> (56,429) <i>argW</i> (66,540)	-	This study	CP088062-3
437022	lc	PT21/28	<i>Stx2a/2a</i>	-	<i>sbcB</i> (56,435) <i>argW</i> (66,543)	-	This study	CP088036-7
437023	lc	PT21/28	<i>Stx2a/2a</i>	-	<i>sbcB</i> (56,427) <i>argW</i> (66,542)	-	This study	CP088034-5
437024	lc	PT21/28	<i>Stx2a/2a</i>	-	<i>sbcB</i> (56,425) <i>argW</i> (66,538)	-	This study	CP088032-3
804533	lc	PT21/28	<i>Stx2a/2a</i>	-	<i>sbcB</i> (56,433) <i>argW</i> (66,414)	-	This study	CP088030-1
811034	lc	PT21/28	<i>Stx2a/2a</i>	-	<i>sbcB</i> (56,438) <i>argW</i> (67,194)	-	This study	CP088028-9
811035	lc	PT21/28	<i>Stx2a/2a</i>	-	<i>sbcB</i> (56,405) <i>argW</i> (66,539)	-	This study	CP088026-7

Table 4.5 – Summary of the sequenced genomes used within this study. Showing their Strain ID, CC11 sub-lineage, Phage Type, *stx* profile, the *stx*-encoding prophage associated with each sample including prophage size and SBI.

4.3.2.2 Comparison of prophage and prophage-like content.

All 23 isolates contained the same number of prophages (n=17) of which two were *stx2a*-encoding prophages (Table 4.4 and Figure 4.6). All prophages in the samples sequenced ranged from 8.2kbp to 144.5kbp in size. Across all 23 samples, 15/17 prophages were considered the same in all samples including prophages 1 (*lexA*), 2 (tRNA-Thr), 3 (*ybhC*), 4 (*yccA*), 6 (*rspA*), 8(*yecA*), 9 (tRNA-Ser), 10 (*ompW*), 11 (*icd*), 12 (*roxA*), 13 (*sbcB* [*stx2a*]), 14 (*yehV*), 15 (*argW* [*stx2a*]), 16 (*argW*) and 17 (*alpA*) (Figure 2 and 4).

There were 2/17 prophages that showed variation, the first type of variation was a single deletion of a 47,389bp region in a large compound prophage 5 (*potC*) in a cattle isolate (432299), different to the isolate exhibiting the LCR (Figure 4.6, Figure 4.7). The deletion event was surrounded by two 5.8kbp homologous sequences containing structural tail proteins and hypothetical genes. The second type of variation observed in prophages 5 (*potC*) and 7 (*yebW*) were related to the 0.65Mbp LCR observed where these prophages acted as the break points and as a result exchanged prophage genes into different locations. (Figure 4.6, Figure 4.7).

Comparing the prophages identified in the outbreak strains to those identified in a well characterised STEC O157:H7 reference genome, strain Sakai (BA000007), it was noted that seven prophages matched between all outbreak genomes and the reference genome. Those prophages included prophage 3 and Sp3 (*ybhC*), prophage 4 and Sp4 (*yccA*), prophage 11 and Sp8 (*icd*), prophage 8 and Sp13 (*yecA*), prophage 9 and Sp14 (tRNA-Ser), prophage 16 and Sp16 (*argW*), prophage 17 and Sp17 (*alpA/ssrA*) (Figure 4.7).

prophages unique to the reference genome only and yellow clusters are prophages unique samples sequenced in this study.

4.3.2.3 Analysis of Shiga toxin-encoding prophages.

All 23 isolates in the outbreak cluster harboured two *stx2a*-encoding prophages (Table 4.5, Figure 4.6 and Figure 4.8); one of which was integrated at the SBI *argW*, while the other was integrated at *sbcB*. The *stx2a*-encoding bacteriophage in sub-lineage 1c is commonly found at *argW*, and phylogenetic analysis showed that this bacteriophage clustered with *stx2a*-encoding phage within sub-lineage 1c (Figure 4.8) ^[170]. The *sbcB* SBI is more commonly occupied by *stx2c*-encoding bacteriophage, and phylogenetic analysis showed that the *stx2a*-encoding bacteriophage integrated at *sbcB* in the outbreak strain was located on a branch mainly comprising of *stx2c*-encoding bacteriophages ^[170].

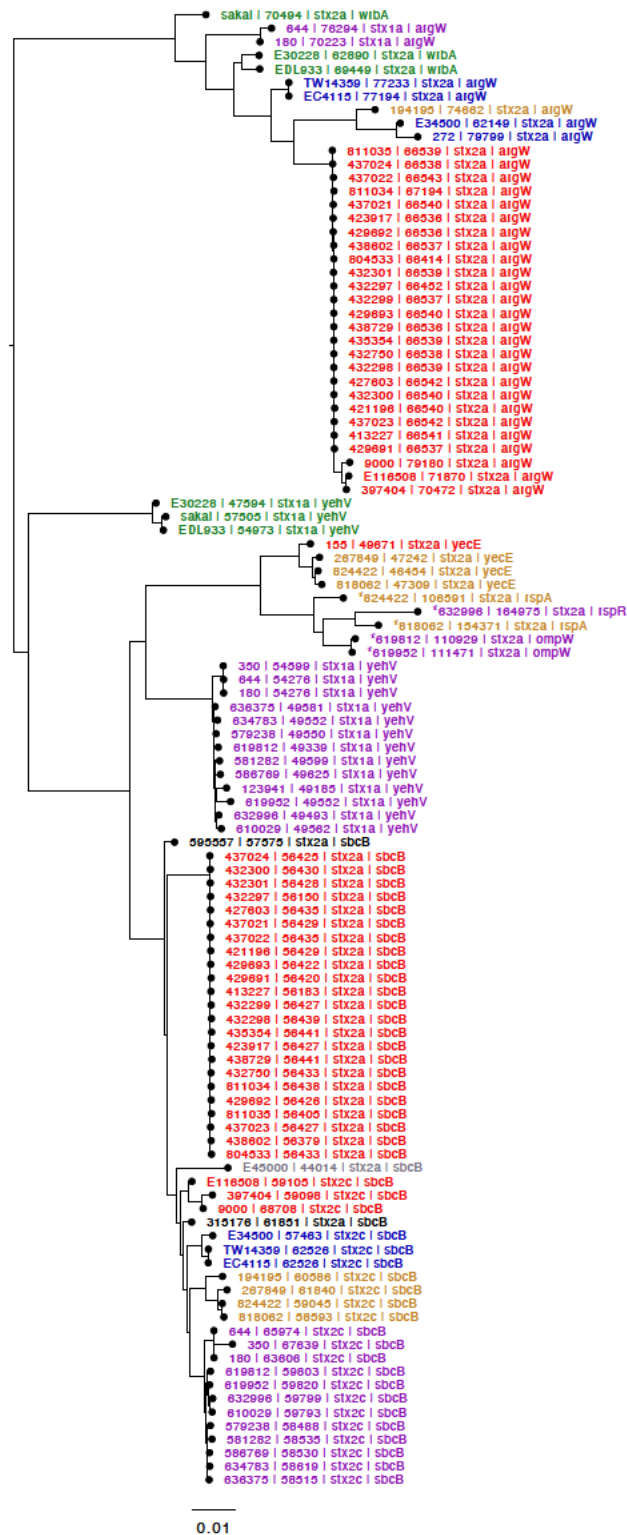


Figure 4.8 – Neighbour joining tree based on Jaccard distances from mash of *stx*-encoding prophages of publicly available samples and the outbreak samples sequenced in this study. Prophages are coloured by sub-lineage of CC11 STEC

O157:H7. Sub-lineage Ia, Green; Ib, Yellow; Ic, Red; I/Ia, Blue; I/Ib, Grey; IIa, Orange; IIb, Black and IIc, Purple.

4.3.2.4 Analysis of plasmid content.

All isolates contained an IncFIB plasmid, the pO157 that is characteristic of STEC O157:H7, ranging in size from 93,689 to 94,050 between the outbreak isolates from 2017, a maximum difference of 361bp and an average size of 93,997bp. In the three temporally distinct isolates from 2019, the size range increased to 95,369bp with a maximum difference of 1,680bp. The approximately 1.6kbp increase between IncFIB plasmids in isolates from 2019 compared to those from 2017, is due to the duplication of insertion elements within those plasmids. One isolate (435354) from 2017, also contained a smaller IncI1- γ plasmid, 85,962bp in size (Table 4.4).

4.3.2.5 Comparison of variant calling methods and phylogeny.

Variant calling of the Illumina sequencing data identified 23 isolates that fell within the same 5-SNP single linkage cluster. The 20 isolates linked to the outbreak in 2017 had a maximum SNP distance of 1 SNP between each other, and a maximum of 5 SNPs when including cases identified in 2019 (Figure 4.9A). Analysis of the Nanopore data identified a maximum of 1 SNP variants between the 20 outbreak isolates and 5 SNPs between all 23 isolates (Figure 4.9B).

The variants identified in the Illumina sequencing data were compared to those identified in the Nanopore data. There were seven discrepant variants called between the two technologies (Table 4.6). Four of the seven mismatches were determined to be false negative variants in the Illumina dataset due to ambiguously

aligning short reads to homologous sequences of the STEC reference genome (Table 4.6). A fifth, this time false positive, variant was called in the Illumina short read data in a single sample due to low coverage over a region of the genome where the variant was called (Table 4.6). The remaining two discrepant variants were deemed to be false negatives in the Nanopore data due to errors associated with calling SNPs within homopolymer sequences (Table 4.6).

When accounting for the above false positive/false negative discrepant variants there remained only a single discrepant variant which was classified as a true variant [Nanopore (G/T reference), Illumina (N/T reference)], accounting for the single SNP difference observed in sample 432299 (Figure 4.9B).

Reference position	Reference base	Variant called	CDS	Encodes	Count of samples with this discrepant variant (out of 23).	Reason for discrepancy (False positive/False negative).
232300	G	A	NON CODING	n/a	1	Low coverage at position from Illumina data (FN)
270579	A	G	R1168G	<i>RhsG</i> core protein with extension	2	Ambiguous alignment at this position from Illumina data (FP)
270595	C	A	T1173N	<i>RhsG</i> core protein with extension	20	Ambiguous alignment at this position from Illumina data (FP)
379516	A	G	NON CODING	n/a	4	Variant missed in homopolymer sequence by Nanopore data (FN)
418442	T	G	E65D	thiogalactoside acetyltransferase	5	Variant missed in homopolymer sequence by Nanopore data (FN)
1681338	C	G	R314G	alanine racemase 2	20	Ambiguous alignment at this position from Illumina data (FP)
1681339	G	C	R314P	alanine racemase 2	20	Ambiguous alignment at this position from Illumina data (FP)

Table 4.6 – The discrepant variant calls between Illumina and Nanopore datasets for all outbreak samples, showing reference position, reference or variant calls, the coding region and associated gene, the count at which this call was observed for samples sequenced in this study and a reason for the observed discrepancy.

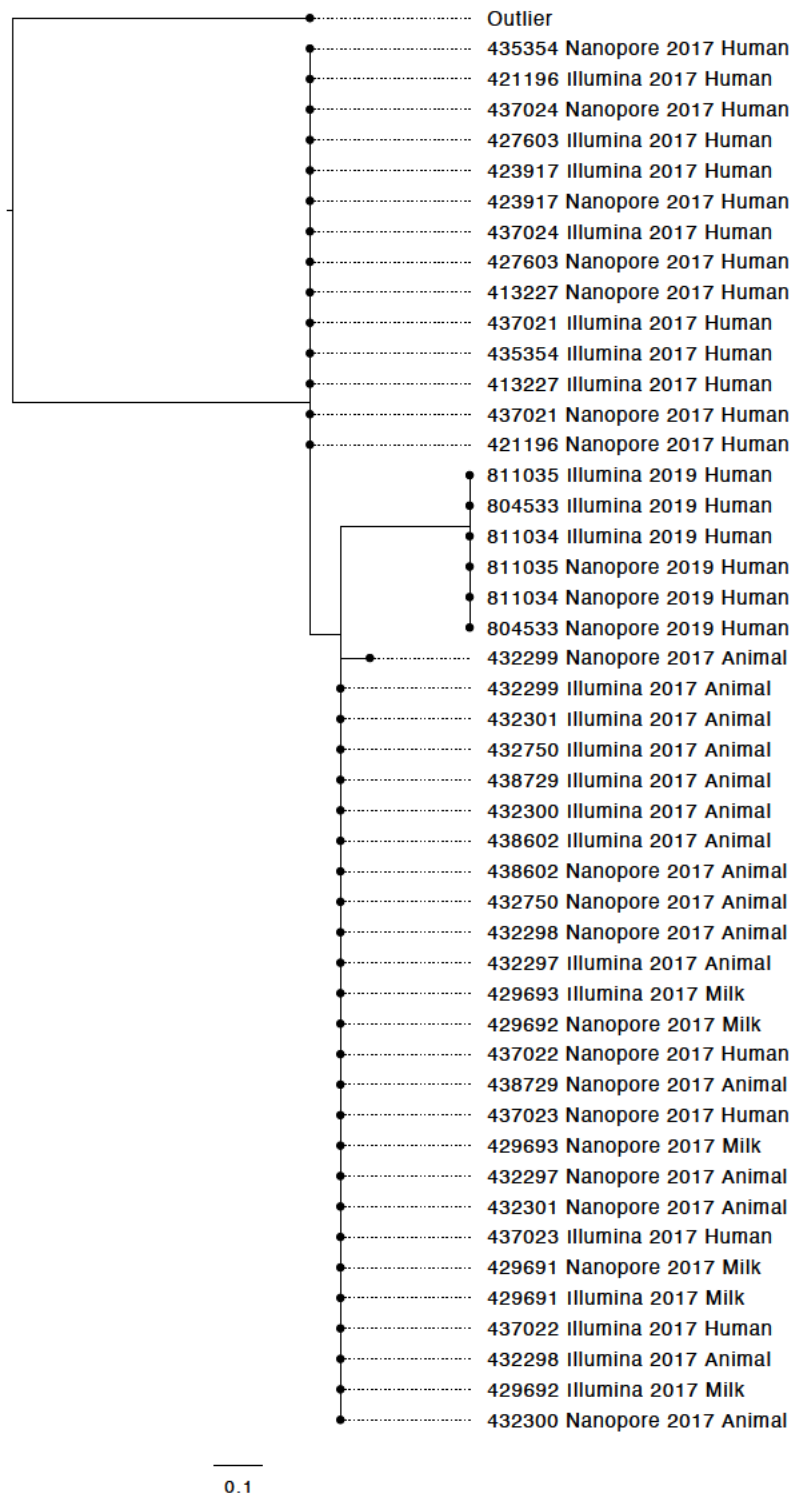


Figure 4.9 –A maximum-likelihood phylogeny showing both Illumina derived and Nanopore derived SNP-typing results for samples sequenced in this outbreak. Annotations include, sample ID, sequencing technology for that sample, year of isolation and sample source.

4.3.3 Outbreak associated with contaminated mud-based obstacle course.

4.3.3.1 Genome features of long-read draft assemblies.

The finalised assemblies of the sequences of the 12 isolates in this outbreak all contained between two and five closed contigs, each supporting a single replicon. The chromosome size of the isolates ranged from 5,480,274bp to 5,530,352bp in length, a maximum difference of 50,078bp and an average size of 5,524,219bp (Table 4.7, Figure 4.10). All isolates contained an 92kbp IncFIB plasmid, the pO157 that is characteristic of STEC O157:H7. Three of the isolates (588888, 588889 and 591229) also contained two smaller plasmids; IncI2 (ranging in size between 57 to 62 kbp), a 63kbp IncX4 plasmid in 588888, and a plasmid 41kbp of unknown Inc type in samples 588889 and 591229 (Table 4.7)

Strain ID	# of contigs	Chromosome size (bp)	# of prophages	Plasmid 1 size (bp) and Inc group	Plasmid 2 size (bp) and Inc group	Plasmid 3 size (bp) and Inc group
391778	2	5,527,979	17	92,954 IncFIB	-	-
401794	2	5,527,760	17	92,958 IncFIB	-	-
588888	4	5,522,628	17	92,686 IncFIB	63,779 IncX4	57,963 IncI2
588889	4	5,527,095	17	92,960 IncFIB	62,479 IncI2	41,465 Unknown Inc
591122	2	5,528,073	17	92,958 IncFIB	-	-
591229	4	5,480,274	17	92,954 IncFIB	62,479 IncI2	41,095 Unknown Inc
591859	2	5,528,066	17	95,400 IncFIB	-	-
591890	2	5,528,673	17	92,959 IncFIB	-	-
592579	2	5,527,957	17	92,958 IncFIB	-	-
592580	2	5,528,456	17	92,956 IncFIB	-	-
595637	2	5,527,567	17	92,955 IncFIB	-	-
595557	2	5,528,386	17	92,951 IncFIB	-	-

Table 4.7 – Summary of the genomes sequenced within this study. Showing strain ID, number of contigs, chromosome size, number of prophages and the size and incompatibility group of plasmids.

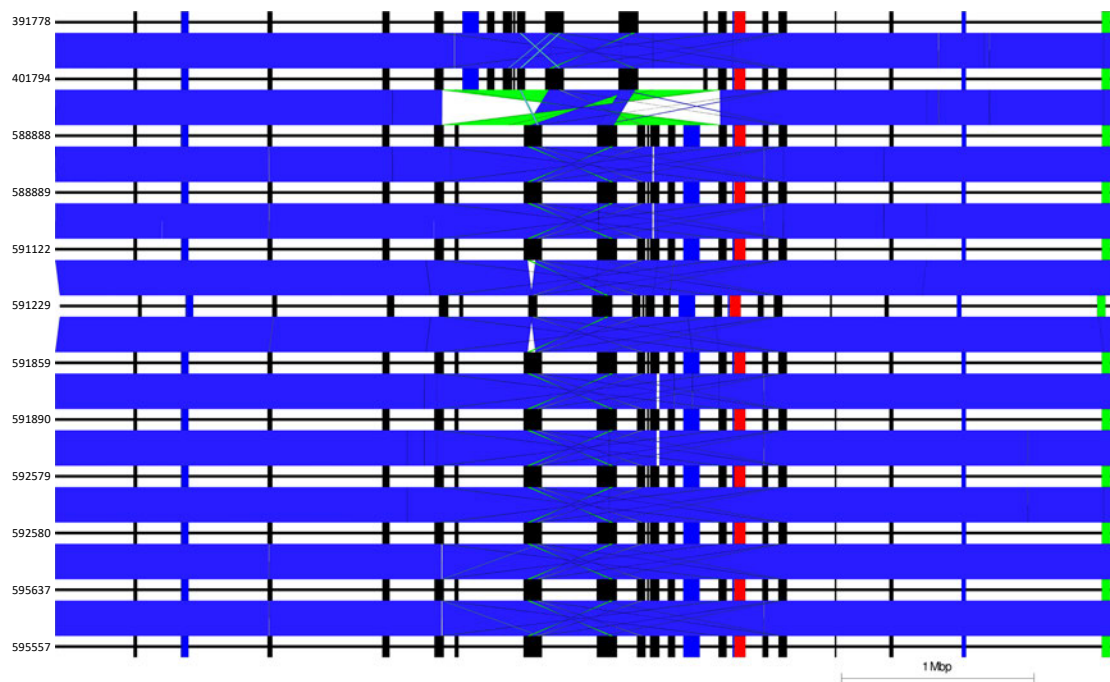


Figure 4.10 – Easyfig plot showing an alignment of linearised chromosomes (horizontal black lines) and loci of prophages (coloured blocks) in all samples in the outbreak in question. Red blocks are *Stx*-encoding prophages, blue blocks are prophage-like regions, green blocks denote the Locus of Enterocyte Effacement and other non-*stx*-encoding prophages are the black blocks. Blue links show regions of sequence similarity, green links show regions of sequence similarity but with different directionality (inversions) and no links denote regions that are different.

4.3.3.2 Analysis of structural variation within outbreak samples.

Aligning and comparing the chromosomes of the isolates within the same 5-SNP single linkage cluster led to the discovery of a large-scale recombination event present in both isolates (391778 and 401794) from the cases identified in 2017, 12 months prior to the outbreak in 2018 (Table 4.8, Figure 4.10). In both isolates (391778 and 401794), the large-scale recombination event is characterised by a 1.468Mbp inversion between prophages 4 (*yccA*) and 12 (tRNA-Ser) (Table 4.8,

Figure 4.10). Additionally, there is an additional 0.452Mbp inversion between prophages 9 (*ompW*) and 10 (*rspR*) in 391778 and 401794 placing this sequence back into the same orientation as the remaining 10 temporally linked isolates (Figure 2).

At either edge of the 1.468Mbp inversion, prophages 4 (*yccA*) and 12 (tRNA-Ser) both have a 12.6kbp homologous sequence spanning from the terminase genes to the host specificity proteins which contains tail, head and portal encoding genes. The 12.6kbp homologous sequences in both prophages 4 (*yccA*) and 12 (tRNA-Ser) share a 99.55% sequence similarity. Similarly, at either edge of the 0.452Mbp inversion, prophages 9 (*ompW*) and 10 (*rspR*) both have a 20kbp homologous sequence, again spanning across the structural genes of the prophage including, head, tail and portal encoding genes. The 20kbp homologous sequences in prophages 9 (*ompW*) and 10 (*rspR*) share a 99.57% sequence similarity.

Taking two exemplar samples, one sample's chromosome in each different chromosomal isoform, 401794 and 588888. There was only a single 588888 read that aligned across the ends of the 12.6kbp homologous sequence in prophage 4 in sample 401794 and similarly only a single read spanning the 12.6kbp homologous sequence in prophage 12. However, there were eight 588888 reads spanning the 20kbp homologous sequence in prophage 8 in sample 401794 and three reads spanning the 20kbp homologous sequence in prophage 12. Both of which denote presence of multiple chromosomal isoforms in each sample read set.

Prophage detected	Gene 5' to prophage in chromosome isoform A	Gene 3' to prophage in chromosome isoform A	Average length (bp) in chromosome isoform A	Order of phage in chromosome isoform A	Relative position in chromosome isoform A (Mbp)	Gene 5' to prophage in chromosome isoform B	Gene 3' to prophage in chromosome isoform V	Average length (bp) in chromosome isoform B	Order of phage in chromosome isoform B	Relative position in chromosome isoform B (Mbp)
1	<i>zur</i>	<i>aphA</i>	19,578	1	0.406-0.425	<i>zur</i>	<i>aphA</i>	19,578	1	0.406-0.425
2 ^a	<i>tRNA-Thr (cgt)</i> ^b	<i>pgrR</i>	26,087	2 ^a	1.103-1.129	<i>tRNA-Thr (cgt)</i> ^b	<i>pgrR</i>	26,087	2	1.103-1.129
3	<i>ybhC</i> ^b	<i>ybhB</i>	38,095	3	1.700-1.738	<i>ybhC</i> ^b	<i>ybhB</i>	38,095	3	1.700-1.738
4	<i>yccA</i> ^b	<i>tRNA-Ser (tga)</i>	48,815	4*	1.970-2.018	<i>yccA</i> ^b	<i>yodB</i>	21,355	4*	1.970-2.020
5	<i>flgK</i>	<i>flgK</i>	38,428	5	2.244-2.282	<i>tRNA-Leu (taa)</i>	<i>yecA</i> ^b	21,355	11	2.075-2.097
6	<i>potC</i> ^b	<i>potB</i>	49,054	6	2.326-2.375	<i>rspA</i>	<i>rspR</i>	96,279	10 ^a	2.434-2.530
7	<i>roxA</i> ^b	<i>phoQ</i>	10,436	7	2.380-2.390	<i>trpA</i>	<i>ompW</i> ^b	106,240	9 ^a	2.814-2.920
8	<i>icd</i> ^b	<i>caeB</i>	43,626	8	2.399-2.443	<i>caeB</i>	<i>icd</i> ^b	43,626	8	3.023-3.067
9	<i>ompW</i> ^b	<i>rspR</i>	98,193	9 ^a	2.546-2.644	<i>phoQ</i>	<i>roxA</i> ^b	10,436	7	3.075-3.086
10	<i>trpA</i>	<i>rspA</i>	104,084	10 ^a	2.927-3.032	<i>potB</i>	<i>potC</i> ^b	49,054	6	3.091-3.139
11	<i>yecA</i> ^b	<i>tRNA-Leu (taa)</i>	21,355	11	3.369-3.390	<i>flgK</i>	<i>flgK</i>	38,428	5	3.182-3.221
12	<i>yodB</i>	<i>tRNA-Ser (cga)</i> ^b	44,436	12*	3.445-3.490	<i>tRNA-Ser (tga)</i>	<i>tRNA-Ser (cga)</i> ^b	42,663	12*	3.446-3.489
13 (<i>stx2a</i>)	<i>yeeA</i>	<i>sbCB</i> ^b	57,681	13 (<i>stx2a</i>)	3.530-3.587	<i>yeeA</i>	<i>sbCB</i> ^b	57,681	13 (<i>stx2a</i>)	3.530-3.587
14	<i>yegQ</i>	<i>yegR</i> ^b	31,007	14	3.675-3.706	<i>yegQ</i>	<i>yegR</i> ^b	31,007	14	3.675-3.706
15	<i>yehV</i>	<i>yehV</i> ^b	44,682	15	3.758-3.803	<i>yehV</i>	<i>yehV</i> ^b	44,682	15	3.758-3.803
16	<i>argW</i> ^b	<i>lacY</i>	8,233	16	4.051-4.060	<i>argW</i> ^b	<i>lacY</i>	8,233	16	4.051-4.060
17	<i>ssrA</i>	<i>alpA</i> ^b	22,154	17	4.334-4.357	<i>ssrA</i>	<i>alpA</i> ^b	22,154	17	4.334-4.357

Table 4.8 – The size, position and integration sites of all 16 prophages within all samples in this study, “a” refers to prophages that appear to be compound prophages (i.e. two or more prophages that are in tandem or one integrated into other, with intact integrase genes). “b” refers to the end in which the Integrase gene (*intA*) is located. “*” denotes prophages on the border of the large

4.3.3.3 Analysis of prophage and prophage-like content.

All 12 isolates contained the same number of prophages (n=17) of which one was a *stx2a*-encoding prophage (Table 4.8, Figure 4.10). All prophages in the samples sequenced ranged from 8kbp to 106kbp in size (Table 4.8). Across all 12 samples, 12/17 prophages were considered the same in all samples including prophages 1 (*zur*), 2 (tRNA-Thr), 3 (*ybhC*), 5 (*flgK*), 6 (*potC*), 8 (*icd*), 11 (*yecA*), 13 (*sbcB*), 14 (*yegR*), 15 (*yehV*), 16 (*argW*) and 17 (*alpA/ssrA*).

There were at least four structural variation events within the prophage content of the nine outbreak isolates. The first was a 48kbp deletion within a terminal compound prophage (prophage 6) in sample 591229 relative to the other eight genomes of the outbreak isolates (Figure 4.10). The second involved deletions in prophage 10 within samples 588889 and 591890. Sample 588889 had a deletion of a prophage anti-repressor gene and sample 5918790 had three deletions, a 1.7kbp deletion including three hypothetical genes and an exonuclease, a two more deletions both regarding individual hypothetical genes (Figures 4.10 and 4.11). The third was a deletion of a single hypothetical protein and a prophage tail measure protein in prophage 4 in sample 595637 (Figures 4.10 and 4.11). The fourth was a 1.2kbp inversion detected in prophage 14 for samples 588889 and 595557 (Figures 4.10 and 4.11).

Additional variation was identified within prophages 4 (*yccA*), 7 (*roxA*), 9 (*ompW*), 10 (*rspR*) and 12 (tRNA-Ser) when the three isolates from the three cases that were geographically but not temporally linked to the outbreak were included in the analysis (Figure 4.10, Figure 4.11). Both prophages 4 (*yccA*) and 12 (tRNA-Ser) were the same in the two isolates from the two cases identified in 2017 (391778 and 401794) and were different when compared to the isolates from the cases identified in 2018

due to the 1.468Mbp chromosomal inversion, described above (Table 3, Figures 1 and 3). Prophages 4 (*yccA*) and 12 (*tRNA-Ser*) are at the breakpoints of the inversion and as a result of homologous recombination, have a exchanged genetic sequences in each chromosomal isoform (Figure 4.10). Prophages 9 (*ompW*) and 10 (*rspR*) in the 2017 isolates (391778 and 401794) were also different compared to the isolates from the cases identified in 2018 due to a 0.452Mbp inversion occurring within homologous sequences within these prophages (Figure 4.10).

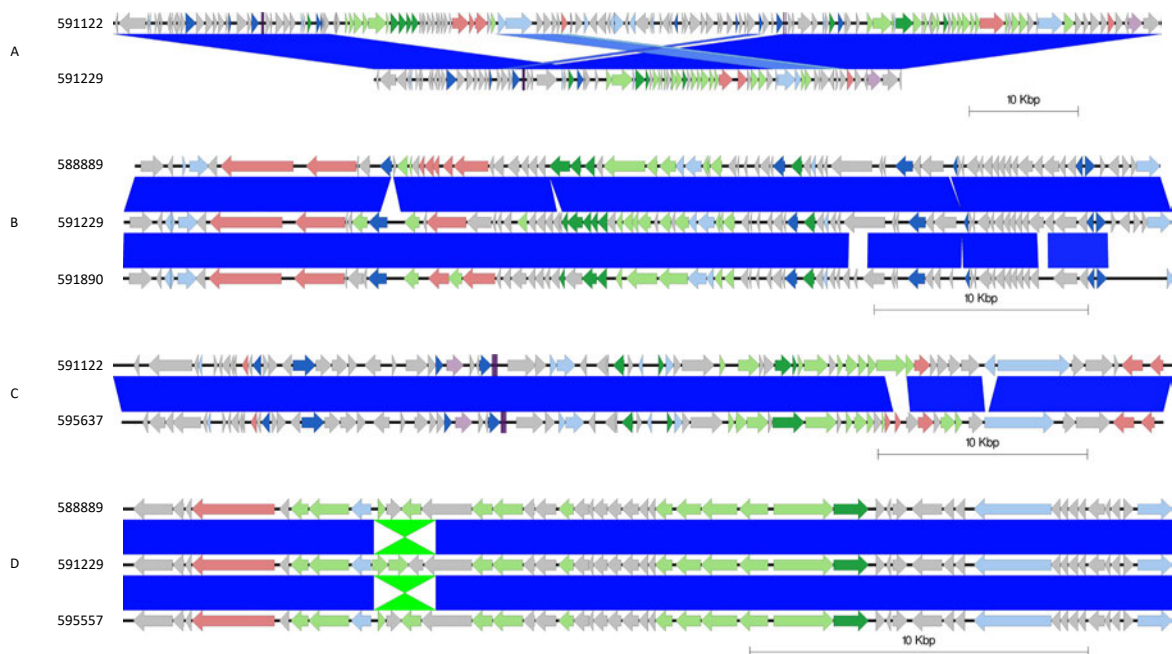


Figure 4.11 – Easyfig plot showing alignments of the four structural variation events within the prophage content in the outbreak samples. Sample 591122 is acting as the reference in each case. Arrows indicate gene directions. *stx* genes are shown in red; recombination/replication genes are shown in light blue; regulation-associated genes are shown in dark blue; effector genes are shown in pink; structure- and lysis-associated genes are shown in light and dark green, respectively; tRNAs are shown as purple lines; finally, hypothetical genes are shown in grey. Blue links show

regions of sequence similarity, green links show regions of sequence similarity but with different directionality (inversions) and no links denote regions that are different.

Comparing the prophages identified in the outbreak strains compared to those identified in a well characterised STEC O157:H7 reference genome, strain Sakai (BA000007) it was noted that seven prophages matched between all outbreak genomes and the reference genome. Those prophages included Sp3 and prophage 3 (*ybhC*); Sp4 and prophage 4 (*yccA*); Sp6 and prophage 6 (*potC*); Sp8 and prophage 8 (*icd*); Sp14 and prophage 12 (tRNA-Ser); Sp16 and prophage 16 (*argW*); Sp17 and prophage 17 (*ssrA/alpA*) (Table 4.8). For Sp11 and recombinant prophage 9 (*ompW/rspA*) (Table 3) across all the samples were the same apart from the modules involved in the large chromosomal re-arrangements (Table 4.8).

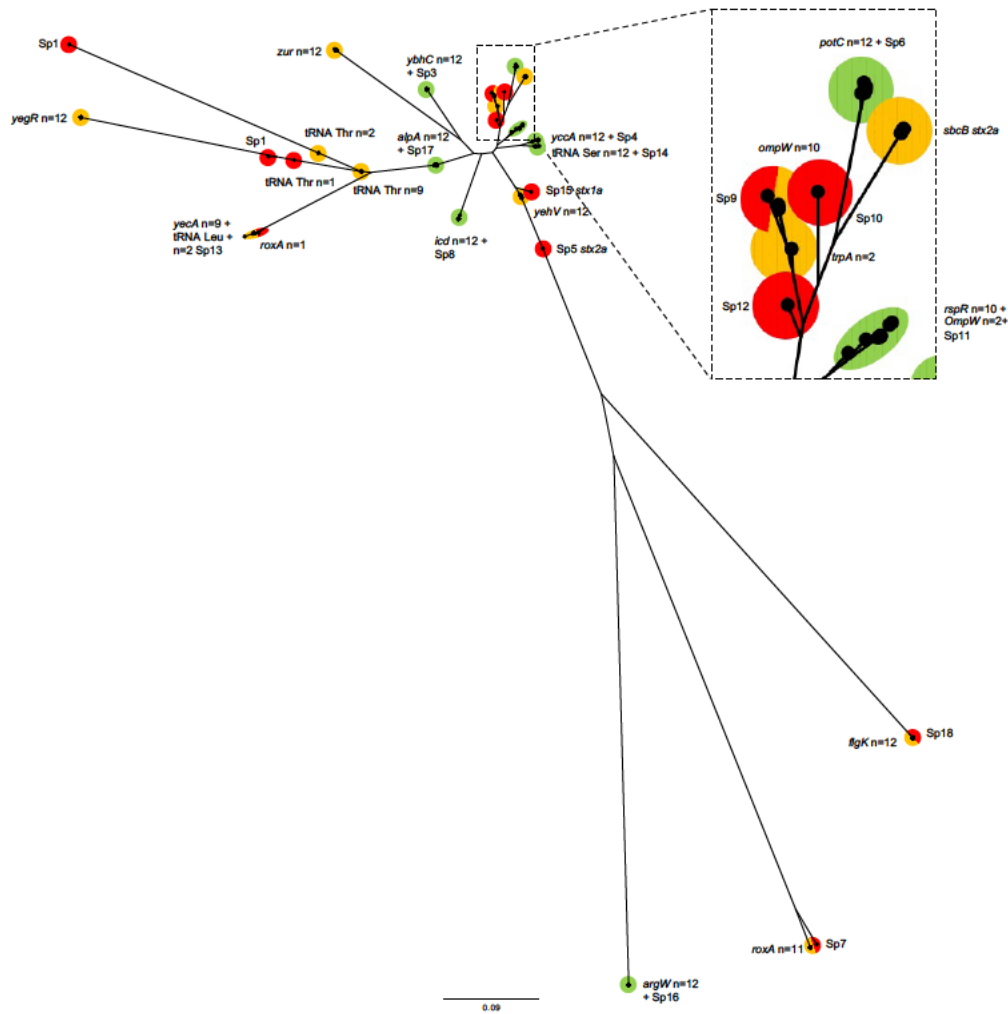


Figure 4.12 – Unrooted neighbour-joining trees of Jaccard distances from mash showing prophages from samples sequenced in this study with prophages from BA000007 (strain Sakai). Prophages clustered in green are prophages shared in samples and reference genome; red clusters are prophages found in the reference genome only and yellow clusters are prophages unique to samples sequenced in this study.

4.3.3.4 Analysis of Shiga toxin-encoding prophages.

All 12 isolates contained a single *stx2a*-encoding prophage, which was integrated at the *stx*-encoding bacteriophage integration site (SBI) *sbcB*, more commonly

associated SBI for *stx2c* (Table 4.8, Figure 4.13, Figure 4.14). Comparing the *stx2a*-encoding prophage from each sample within the outbreak showed no variation. We also compared an outbreak example *stx2a*-encoding prophage with *stx2a*-encoding prophage a PT49 STEC O157:H7 (E45000) from a different sub-lineage (I/IIb) but sharing the same SBI (*sbcB*) and a PT8 STEC O157:H7 (315176) from the same sub-lineage (IIb) also harbouring *stx2a*-encoding prophage at the SBI (*sbcB*). Compared to E45000 the outbreak *stx2a*-encoding prophage showed similarity across the *nin*, lysis, *stx* and structural genes of the prophage with only variation appearing with site-to-site recombination and host specificity genes. When compared to 315176 showed more similarity to the outbreak *stx2a*-encoding prophage with only two transposase related genes difference (Table 4.8, Figure 4.13, Figure 4.14).

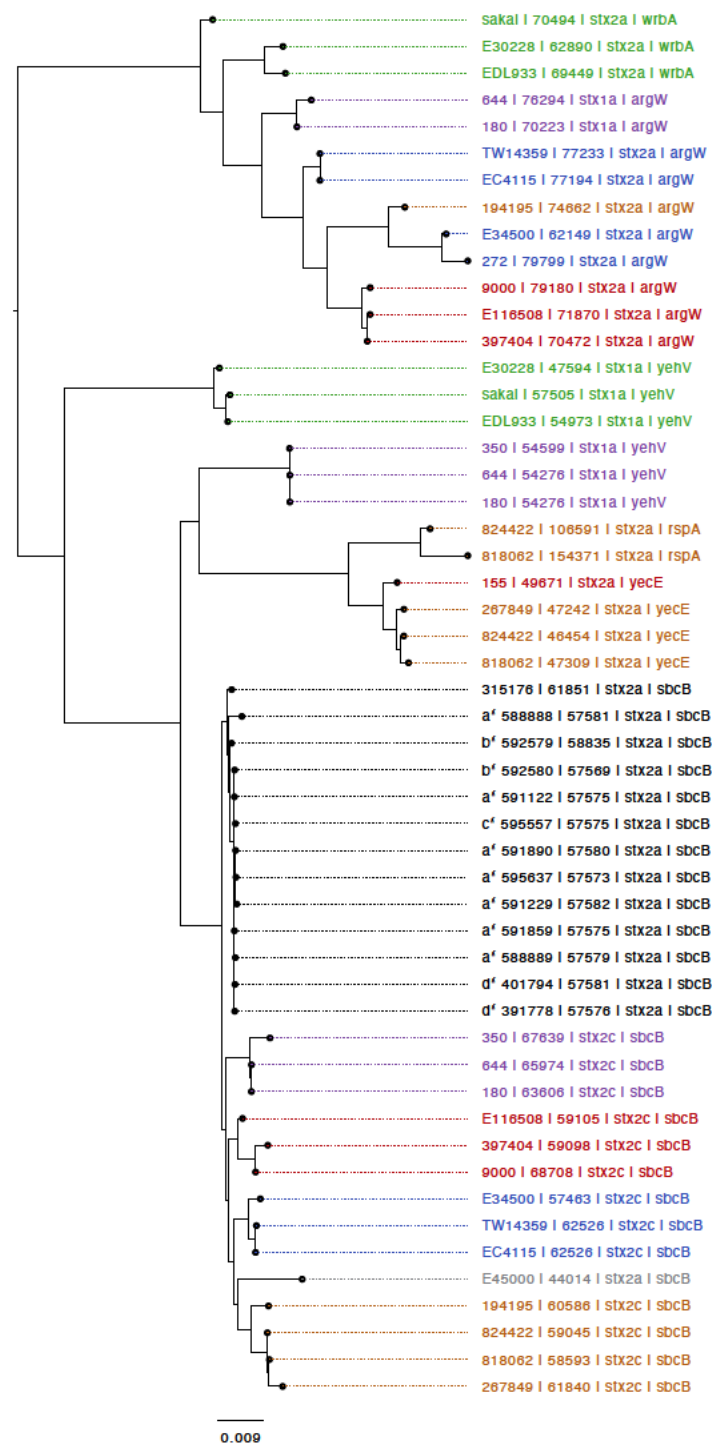


Figure 4.13 – Neighbour joining tree based on Jaccard distances from mash of *stx*-encoding prophages from publicly available samples and from the outbreak samples sequenced in this study (labelled with * preceding strain ID). Prophages are coloured by sub-lineage of STEC O157:H7. Sub-lineage Ia, Green; Ib, Yellow; Ic, Red; I/IIa, Blue; I/IIb, Grey; IIa, Orange; IIb, Black and IIc, Purple. Outbreak samples'

annotations, a, race participant; b, household case; c, geographically distinct and isolated in 2018; d, geographically distinct and isolated in 2017.



Figure 4.14 – Easyfig plot showing an alignment of an outbreak exemplar (591229) *stx2a*-encoding prophage comparing to *stx2a*-encoding prophages which share the *sbcB* SBI from STEC CC11 sub-lineages IIa (E45000) and IIb (315176). Arrows indicate gene directions. *stx* genes are shown in red; recombination/replication genes are shown in light blue; regulation-associated genes are shown in dark blue; effector genes are shown in pink; structure- and lysis-associated genes are shown in light and dark green, respectively; tRNAs are shown as purple lines; finally, hypothetical genes are shown in grey. Blue links show regions of sequence similarity, green links show regions of sequence similarity but with different directionality (inversions) and no links denote regions that are different.

4.3.3.5 Prophage-like regions and pathogenicity island variation.

In all 12 isolates the prophage-like regions were the same size and in the same SBI for each respective prophage-like region. This included the locus of enterocyte effacement (LEE), which is labelled as PLE6 in the outbreak samples and is located at *se/C* (Table 4.9) and is the typical location for the LEE across STEC O157:H7. The only difference to note is the location of PLE3 which varies between two samples (391778 and 401794) due to PLE3 being contained within the observed

large chromosomal rearrangement between the genomes (Figure 4.10 and Table 4.9) and thus is inverted relative to the other chromosomes. Comparing the LEE from the outbreak genomes to strain Sakai's LEE with only a single gene different, a IS629 transposase (*tnpA*) gene found in the outbreak sample's LEE and not strain Sakai's LEE (Table 4.9).

4.3.3.6 Analysis of plasmid content.

The IncFIB pO157 plasmids from all 12 isolates ranged from 92,686bp – 95,400bp in size. The gene content remained similar across all samples except for sample 591859 that had acquired an extra 2kbp integrative element containing hypothetical genes (Table 4.7). The two 62kbp IncI2 plasmids were the same apart from the synteny of a single hypothetical gene and a tyrosine recombinase (*xerC*) gene. The smaller 57kbp plasmid has lost *tra* conjugation cassette, transposases and hypothetical genes compared to the other two 62kbp plasmids. The content of the IncI2 plasmids included a *tra* conjugation cassette and a type IV secretion system (Table 4.7). The two plasmids of unknown Inc type (41kbp) were found to be identical in terms of content except a single missing hypothetical gene in the plasmid from sample 591229 relative to the plasmid from sample 588889. The content of these plasmids also included a *tra* conjugation cassette and a type IV secretion system. The single 36kbp IncX4 plasmid from sample 588888 was not found in any other samples.

Prophage-like region detected	Gene 5' to PLE in chromosome isoform A	Gene 3' to PLE in chromosome isoform A	Average length (bp) in chromosome isoform A	Relative position in chromosome isoform A (Mbp)	Gene 5' to PLE in chromosome isoform B	Gene 3' to PLE in chromosome isoform V	Average length (bp) in chromosome isoform B	Relative position in chromosome isoform B (Mbp)
PLE1	Hypothetical	<i>intA</i>	6,085	0.652-0.659	Hypothetical	<i>intA</i>	6,085	0.652-0.659
PLE2	Hypothetical	<i>yjhS</i>	33,706	0.659-0.693	Hypothetical	<i>yjhS</i>	33,706	0.659-0.693
PLE3	<i>ycdU</i>	<i>tRNA-Ser (gga)</i>	85,934	2.116-2.202	<i>tRNA-Ser (gga)</i>	<i>ycdU</i>	85,934	3.266-3.352
PLE4	Hypothetical	<i>yeeA</i>	8,597	3.519-3.528	Hypothetical	<i>yeeA</i>	8,597	3.519-3.528
PLE5	<i>flgK</i>	<i>pitB</i>	23,334	4.710-4.733	<i>flgK</i>	<i>pitB</i>	23,334	4.710-4.733
PLE6	<i>selC</i>	<i>yicL</i>	44,010	5.438-5.483	<i>selC</i>	<i>yicL</i>	44,010	5.438-5.483

Table 4.9 – The size, position and location of any prophage-like elements within all outbreak samples in this study.

Reference pos	Reference base	Variant called	CDS	Encodes	Count of samples with this discrepant variant.
270,595	C	A	T1173N	RhsG core protein with extension	1
379,516	A	G	NON CODING	NON CODING	3
551,747	T	G	None	None	1
648,904	T	C	SYNONYMOUS	hypothetical protein	1
1,665,324	T	G	V212G	hypothetical protein	3
1,681,338	C	G	R314G	alanine racemase 2	12
1,681,339	G	C	R314P	alanine racemase 2	12
1,859,733	T	G	NON CODING	NON CODING	1
2,419,218	A	G	SYNONYMOUS	PEP-dependent phosphotransferase enzyme III for cellobiose	8
3,106,426	A	C	V17G	o-succinylbenzoyl-CoA synthase	10
3,449,117	G	C	NON CODING	NON CODING	2
4,086,858	T	C	SYNONYMOUS	putative ATP-binding component of a transport system	3
4,370,590	T	G	NON CODING	NON CODING	1
4,701,114	C	T	SYNONYMOUS	putative outer membrane usher protein precursor	6
4,955,915	G	T	SYNONYMOUS	membrane protein	1
5,282,356	C	T	None	None	1

Table 4.10 – The discrepant variant calls between Illumina and Nanopore datasets during SNP typing for all 12 outbreak samples.

Showing reference base and associated variant call, any gene annotation and the count of a given variant across all 12 samples.

4.3.3.7 Comparison of variant calling methods and phylogeny.

Analysis of the Illumina sequencing data showed that the 12 isolates that fell within the 5 SNP single linkage cluster had a maximum SNP distance of 6 SNPs between all samples (Figure 4.15). Analysis of the Nanopore data identified a maximum of 23 variants between the 12 isolates. Of these 23 variants, 14 were associated with only two of the 12 isolates and located within a putative transposase (reference genome locus tag – ECs_5183).

The variants identified in the Illumina sequencing data were compared to those identified in the Nanopore data. The six variants called in the Illumina dataset were all present in the Nanopore dataset and there were 16 variants in the Nanopore data that were not identified in the Illumina data (Table 4.10). Two variants both positioned in a known paralogous sequence (alanine racemase) were discrepant in all 12 sequence alignments (Table 4.10). There were three discrepant variants that were detected in between six and 10 isolates, and 11 that were present in one to three isolates (Table 4.10).

Once the false positive variants in the Nanopore data were identified for each sequence and masked in the phylogeny, as described previously ^[171], eight samples had zero variants called between both Nanopore and Illumina data and were placed in the same position in the phylogenetic tree (Figure 4.15). True variants between the two technologies were detected for four samples. Analysis of the Nanopore data for two samples (592580 and 595637), revealed a single variant with the Illumina counterpart, and two samples (591229 and 588889) were >10 SNPs from their Illumina counterparts (Figure 4.15), and this was due to the 14 variants were positioned in the putative transposase ECs_5183 in the reference genome,

described above, that were masked during the analysis of the Illumina data because the bases were not called and produced N calls.

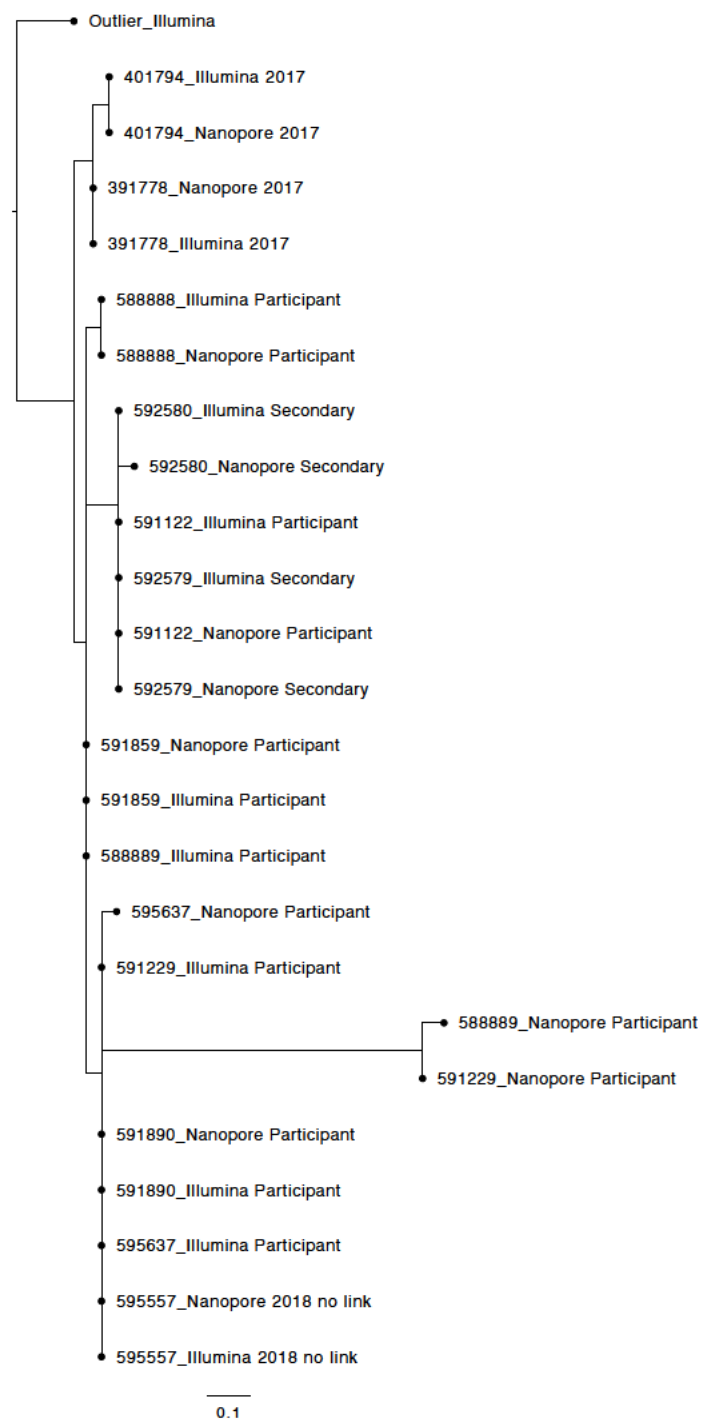


Figure 4.15 – Maximum-likelihood phylogeny showing both Illumina derived and Nanopore derived SNP-typing results for samples sequenced in this outbreak. Annotations detail, sample ID, the sequencing technology used and the case details.

4.3.4 Outbreak associated with animal contact during a lambing event.

4.3.4.1 Genome features of long-read draft assemblies.

Of the 26 samples sequenced in this outbreak set, complete closed genomes were generated for ten genomes which contained either two or three closed circular contigs, each harbouring a single complete replicon. Chromosome size of the genomes ranged from 5,357,939bp to 5,460,589bp in length with a maximum difference of 102,650bp and an average of 5,401,488bp (Table 4.11, Figure 4.16).

All isolates contained a 90-91kbp IncFIB plasmid, the pO157 that is characteristic of STEC O157:H7. A single isolate (60202) contained an extra plasmid of unknown replicon type.

Strain ID	# of contigs	Chromosome size (bp)	Total genome size (bp)	Plasmid 1 size (bp) and Inc group	Plasmid 2 size (bp) and Inc group	# of prophages
60188	2	5,357,975	5,449,416	91,441 (IncFIB)	-	15
60189	2	5,415,603	5,505,913	90,310 (IncFIB)	-	16
60190	2	5,358,431	5,449,882	91,451 (IncFIB)	-	15
60192	2	5,417,625	5,509,073	91,448 (IncFIB)	-	16
60197	2	5,358,015	5,449,464	91,449 (IncFIB)	-	15
60198	2	5,411,622	5,503,072	91,450 (IncFIB)	-	16
60202	3	5,417,503	5,579,030	91,450 (IncFIB)	70,077 (Inc unknown)	15
60203	2	5,357,939	5,449,396	91,457 (IncFIB)	-	15
60207	2	5,459,573	5,551,016	91,443 (IncFIB)	-	16
60209	2	5,460,589	5,552,038	91,449 (IncFIB)	-	16

Table 4.11 – Summary of the genomes sequenced within this outbreak. Showing strain ID, number of contigs, chromosome size, number of prophages and the size and incompatibility group of plasmids.

4.3.4.2 Analysis of structural variation within outbreak samples.

Aligning and comparing the chromosomes of the isolates within the same 5 SNP single linkage cluster led to the discovery of several large-scale recombination events. Samples 60188, 60190 and 60197 all shared the same synteny and location of all prophage content (n=15). Relative to the above samples, sample 60203 also shared the same prophage location (n=15) however there was a 163.7kbp inversion that has occurred between two identical sets of 1.2kbp insertion elements (IS629/IS25) on the chromosome (Figure 4.16, Table 4.11) approximately 4.5Mbp from *oriC*. Also relative to 60188, 60190 and 60197, two samples 60189 and 60192 have the same chromosomal synteny with an additional 60.2kbp prophage (totalling n=16) located at *ompW*. Relative to 60189 and 60192, samples 60207 and 60209 shared the same chromosomal synteny however the prophage located at *rspA* was 42kbp larger in the later samples due to the integration of a prophage in tandem with the prophage already residing at *rspA*.

The remaining two samples (90198 and 90202) showed the least chromosomal synteny relative to other samples. Sample 60198 shared the most similarity to samples 60189 and 60192 however, there was a large 309kbp chromosomal inversion between the prophages located at *ompW* and *rspA* which also spans the terminus of the chromosome. At either edge of this inversion, both prophages have a 11kbp homologous sequence spanning across the structural genes of the prophage head, tail, portal and host specific encoding genes (Figure 4.16).

Sample 60202 is most closely to the structure of the chromosomes of 60188, 60190 and 60197. 60202 has a 248.6kbp inverted translocation of the chromosome involving the mobilisation of two prophages. The breakpoints of the element were

formed between the prophage located at *potC* and the *stx2c*-encoding prophage (Figure 4.16).

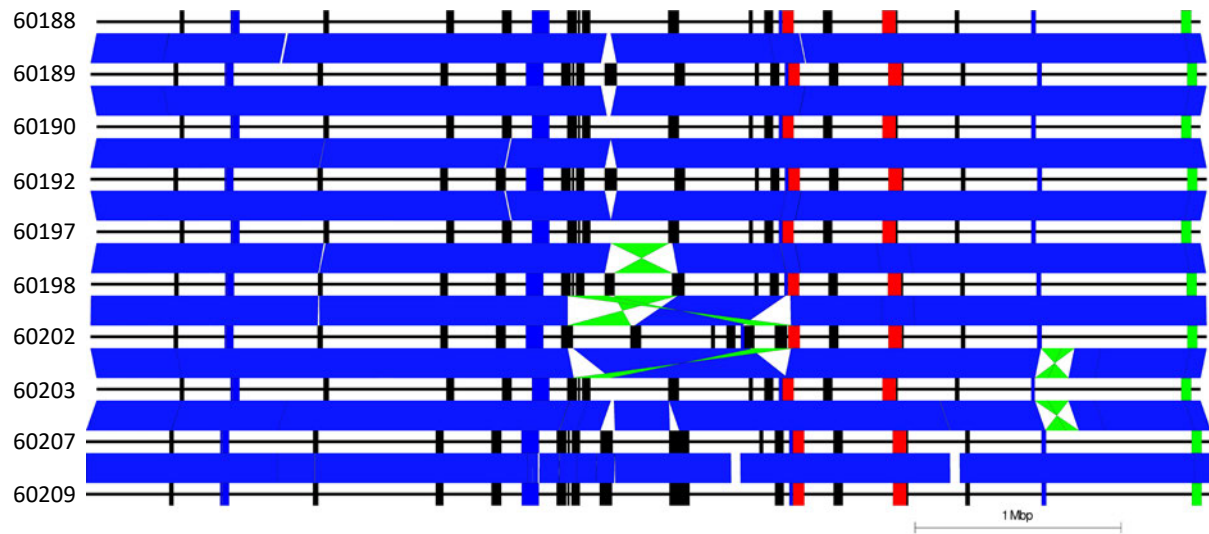


Figure 4.16 – Easyfig plot showing an alignment of linearised chromosomes and loci of prophages in all samples in the outbreak in question. *Stx*-encoding prophages are highlighted as red blocks; blue blocks for prophage-like regions; green blocks for the Locus of Enterocyte Effacement (LEE) and black blocks for other non-*stx*-encoding prophages. Blue links show regions of sequence similarity, green links show regions of sequence similarity but with different directionality (inversions) and no links denote regions that are different.

4.3.4.3 Analysis of prophage and prophage-like content.

For the 10 genomes that generated complete assemblies, the number of prophages across the samples varied between 15 and 16 prophages (Table 4.11). The additional prophage to all samples harbouring 15 prophages was located at *ompW*. There is a compound prophage located at *rspA* which in two samples (60207 and 60209) had an additional prophage integrated sequentially to the prophage, however

this was not counted as there was no chromosomal sequence between the prophages (Figure 4.16). In total there were eight prophages that were considered to be the same prophage across all 10 chromosomes analysed. These prophages included prophages located at *lexA*, tRNA-Thr, *ybhC*, *yehV*, tRNA-Ser, *argW*, *argW* (*stx2a*) and *alpA/ssrA* (Figure 4.17).

There were two large recombination events where the endpoints were located in prophages regions which increased the prophage diversity observed. Firstly, an inverted translocation affecting the prophage located at *potC* and the *stx2c*-encoding prophage in sample 60202 (Figure 4.16, Figure 4.17). Secondly, an inversion between *ompW* and *rspA* in sample 60198 (Figure 4.16, Figure 4.17). This had the effect of exchanging modules of prophages and thus increasing the diversity observed in each of the strains mentioned above (Figure 4.16).

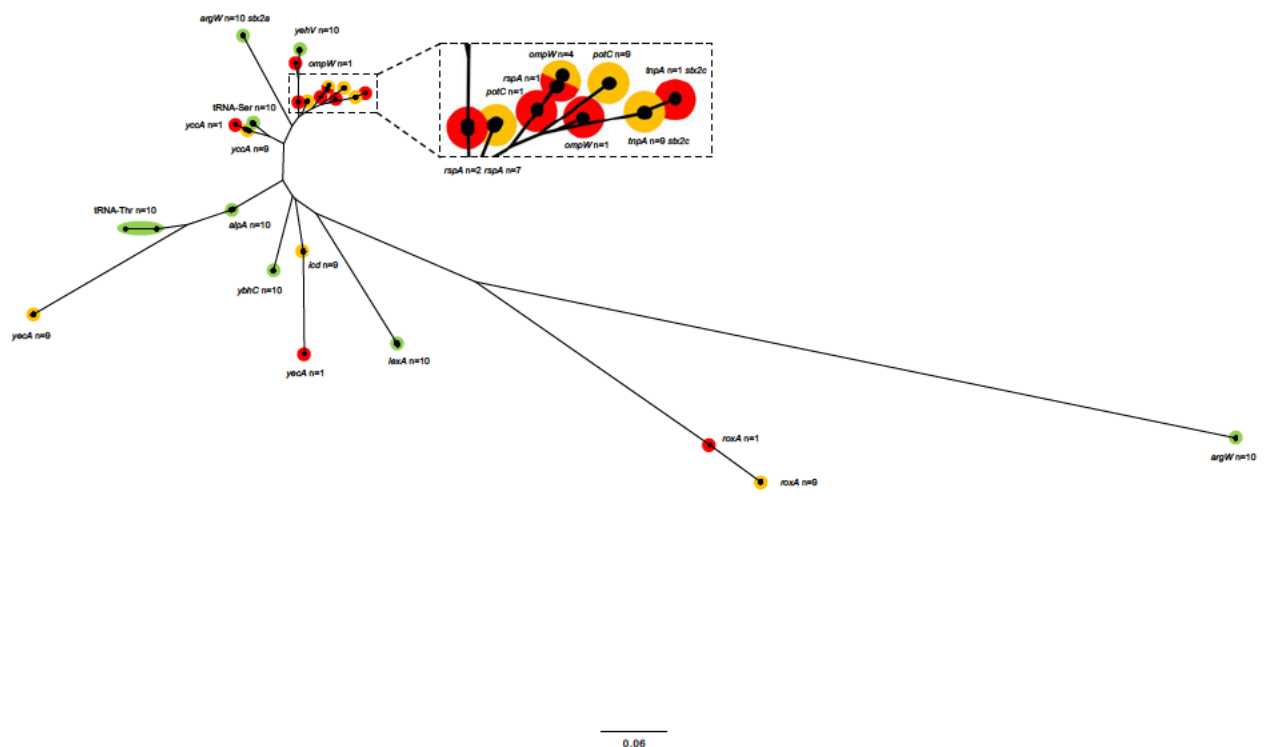


Figure 4.17 – Unrooted neighbour-joining tree of Jaccard distances from mash showing prophages from samples sequenced in this study. Prophages clustered by green are prophages shared in all samples; red clusters are unique prophages (only found in a single sample) and yellow clusters are prophages which are shared by some but not all of samples.

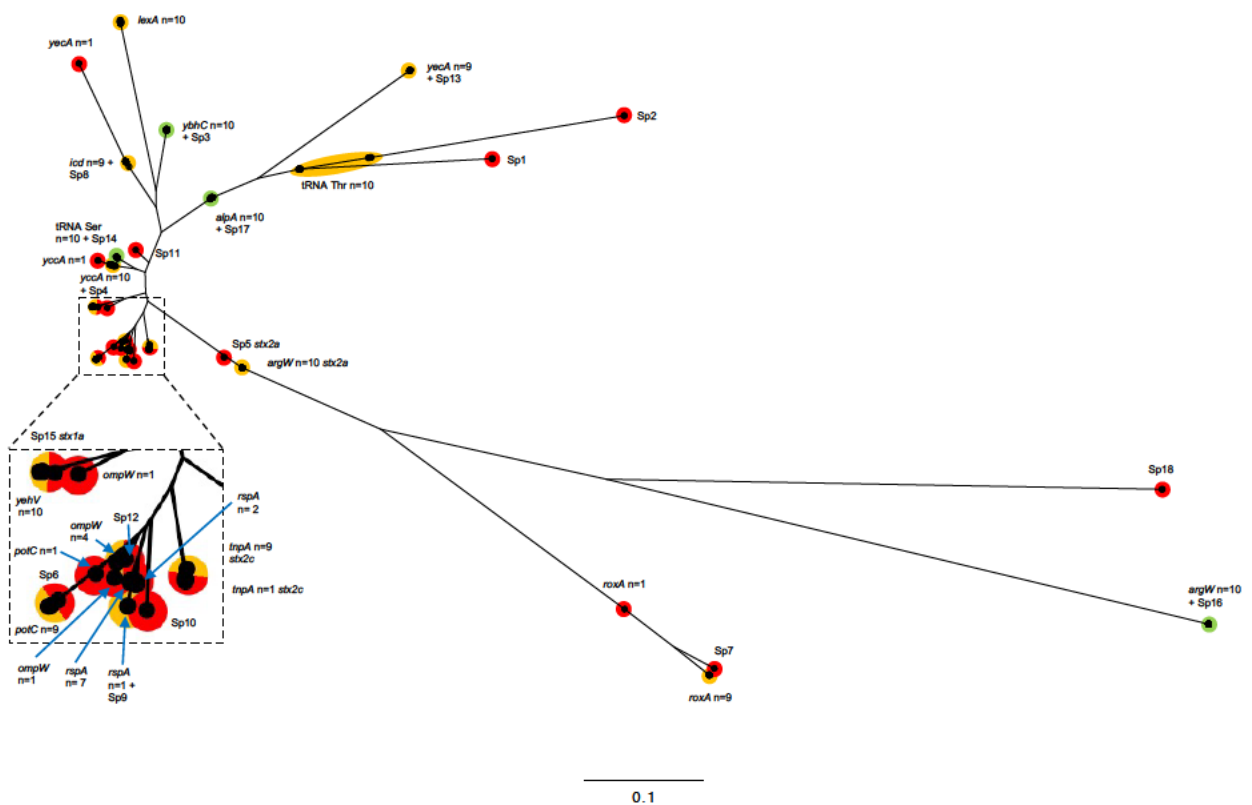


Figure 4.18 – Unrooted neighbour-joining trees of Jaccard distances from mash showing prophages from samples sequenced in this study with prophages from BA000007 (strain Sakai). Prophages clustered in green are prophages shared in samples and reference genome; red clusters are prophages found in the reference genome only and yellow clusters are prophages that are unique from samples sequenced in this study.

Comparing the prophages identified in the outbreak strains compared to those identified in a well characterised STEC O157:H7 reference genome, strain Sakai (BA000007) it was noted that only four prophages matched between all outbreak genomes and the reference genome highlighting the dynamic prophage content of this outbreak strain. Those prophages included Sp3 and the prophage located at *ybhC*; Sp14 and the prophage located at tRNA-Ser; Sp16 and the prophage located at *argW*; Sp17 and the prophage located at *ssrA/alpA* (Figure 4.18).

4.3.4.4 Analysis of Shiga toxin-encoding prophages.

All 10 assembled samples contained a *stx2a*-encoding prophage and a *stx2c*-encoding prophage. The *stx*-encoding bacteriophage integration site (SBI) of both the *stx2a*-encoding and *stx2c*-encoding prophage, *argW* and *sbcB* respectively, were typical of what is observed in STEC O157:H7 (Figure 4.18).

When comparing the *stx*-encoding prophages from these outbreak strains to others from the STEC O157:H7 prophage population, the sub-lineage of the outbreak strains (Ic) both the *stx2a*-encoding and *stx2c*-encoding prophages clustered with those also associated with sub-lineage Ic (Figure 4.19).

The *stx2c*-encoding prophages appeared to have integrated at a transposase gene which sits adjacent to the *sbcB* gene, which is atypical of *stx2c*-encoding prophages in general and within sub-lineage Ic. This could suggest a form of bacteriophage scarring, where a fragment of a previous prophage occupier of this SBI remained behind resulting in the observed extra transposase gene.

Comparing the *stx2a*-encoding prophages, there was no diversity observed in terms of larger structure or low-level variation (Figure 4.19). The *stx2c*-encoding prophages

were also identical apart from sample 60202, which had a section/module different compared to the other sample's *stx2c*-encoding prophage. This is due to the inverted translocation that is observed in this sample, where one end of this region occurs in the *stx2c*-encoding prophage of this sample and the other in a non-*stx*-encoding prophage. (Figure 4.16, Figure 4.19)

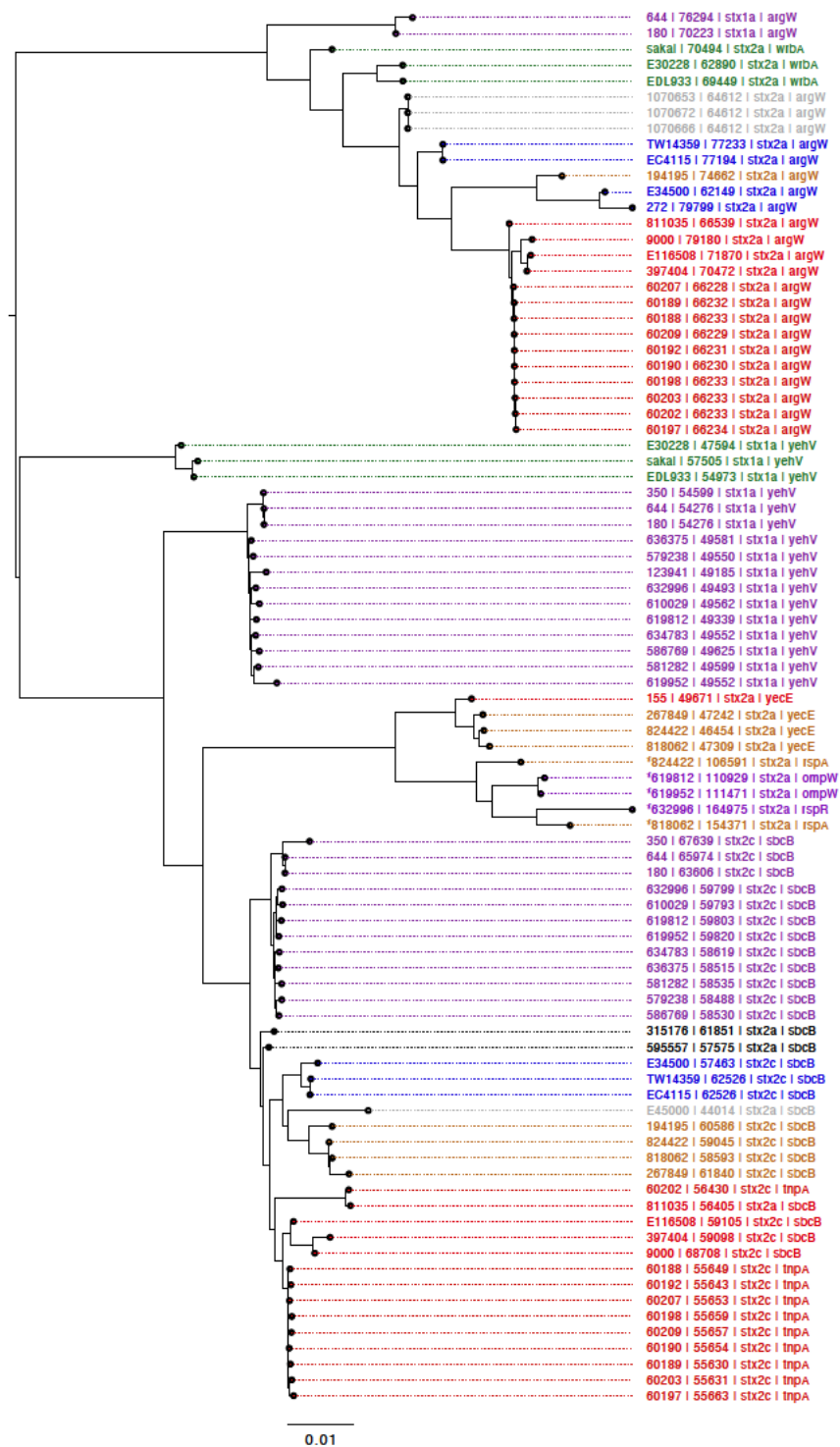


Figure 4.19 – Neighbour joining tree based on Jaccard distances from mash of *stx*-encoding prophages of publicly available samples and the outbreak samples sequenced in this study (labelled with * preceding strain ID). Prophages are coloured by sub-lineage of STEC O157:H7. Sub-lineage Ia, Green; Ib, Yellow; Ic, Red; I/IIa, Blue; I/IIb, Grey; IIa, Orange; IIb, Black and IIc, Purple.

4.3.4.5 Analysis of plasmid content.

The IncFIB pO157 plasmids from all 10 genomes that resulted in a complete assembly, ranged from 90,310bp – 91,457bp in size with an average of 91,335bp. The pO157 in sample 60189 was shorter than the remaining samples by 1.1kbp due to a truncated haemolysin gene (*exhA*), otherwise the remaining gene content remained the same (Table 4.11).

4.3.4.6 Comparison of variant calling methods and phylogeny.

Analysing the Illumina sequencing data showed that the 26 genomes that fell within the 5-SNP single linkage cluster had a maximum SNP distance of 4 SNPs between all samples (Figure 4.20). When including the Nanopore data, a maximum of 4 variants remained between the isolates. Out of the 26 pairs of data (Illumina generated and Nanopore generated), there were only four samples that showed variation, those being samples 60194, 60198, 60199 and 60208. All four instances had a single SNP difference between the two sets of sequencing data (Figure 4.20). The SNP difference for samples 60194, 60198 and 60208 all were determined to be false negative calls on the Illumina dataset failing to call a genuine SNP position within a paralogous gene (*rshG*) where the Nanopore equivalent correctly called the variant (Table 4.12). Finally, sample 60199 had a false positive variant call on a different position within the same paralogous *rshG* gene in the Illumina dataset due to ambiguous alignment over that position.

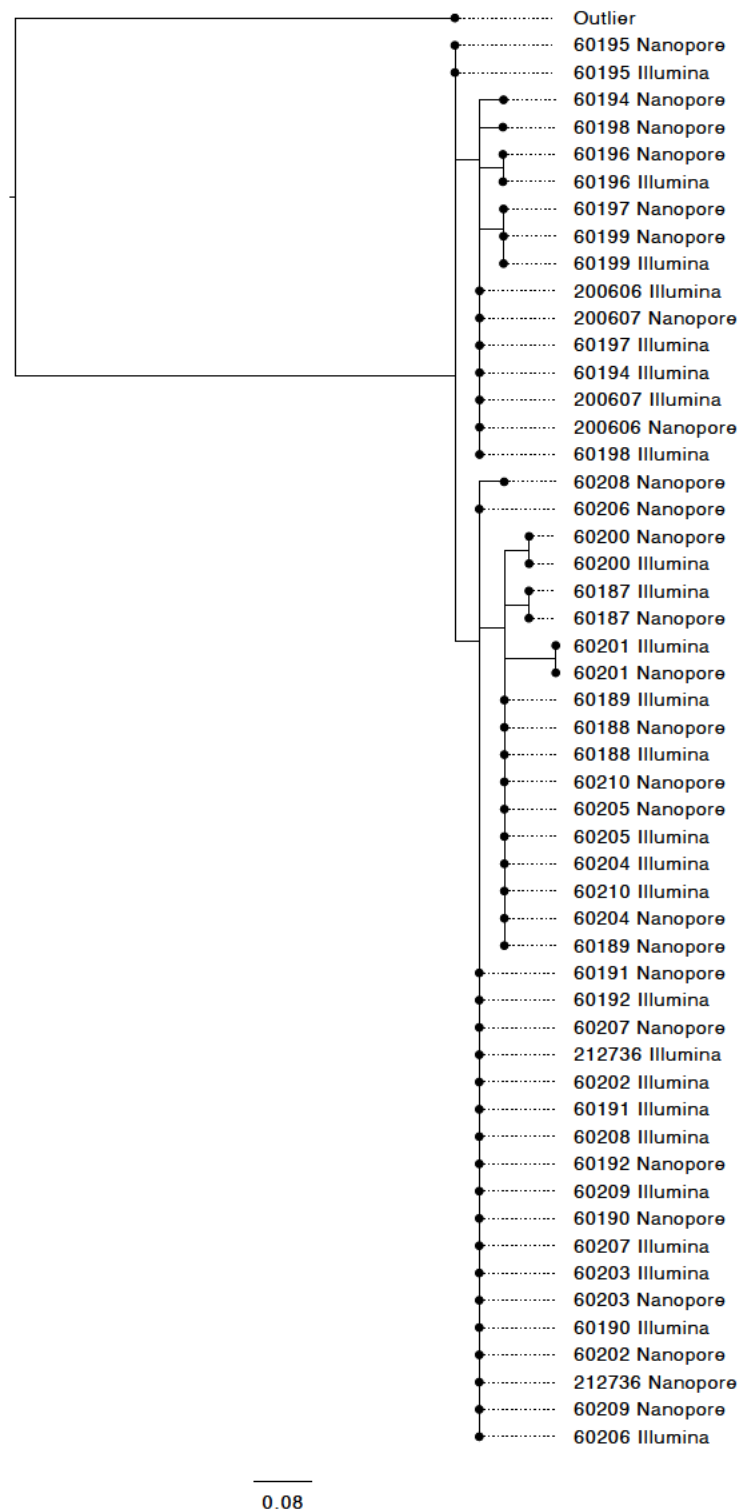


Figure 4.20 – Maximum-likelihood phylogeny showing both Illumina derived and Nanopore derived SNP-typing results for samples sequenced in this outbreak.

Reference pos	Variant called	CDS	Encodes	Count of samples with this discrepant variant.
270579	G	R1168G	RhsG core protein with extension	6
270595	A	T1173N	RhsG core protein with extension	24
379516	G	NON CODING		3
1072231	T	None	None	1
1681359	T	G321*	alanine racemase 2	1
3081442	G	None	None	1
4184453	G	None	None	1
5114397	T	SYNONYMOUS	hypothetical protein	2
5171593	T	S161T	phosphonate metabolism	1

Table 4.12 – The discrepant variant calls between Illumina and Nanopore datasets for SNP typing for all outbreak samples.

4.4. Chapter 4 discussion.

4.4.1 Comparison of variant calling methods and phylogeny.

The SNP typing analyses from each outbreak showed systemic differences associated with each technology that was used to sequence the strains. Specifically, the basecalling errors related to homopolymer detection observed in Nanopore data [200,201]. However, more of the observed errors were due to ambiguous alignment of Illumina (short) reads to homologous and paralogous regions of the STEC genome [202], in particular prophages. These were then miscalled during the variant calling process where individual positions/nucleotides that have ambiguous reads aligned to them are considered mixed due to the lack of consensus within the alignment. However, in reality, they are not mixed, the issue is the alignment algorithms incorrectly align reads to these homologous locations. Further evidence of this effect was observed when aligning Illumina (short) reads to paralogous genes within the strain Sakai reference genome where the majority of samples across all outbreaks found discrepant variant calls within the *RhsG* and *aln* genes. This comparison also demonstrates the importance of masking these regions within the reference genome to produce accurate and meaningful results between the two technologies [200].

Despite the above, the comparison of the Nanopore sequencing data with the Illumina sequencing data confirmed the close genetic relatedness of the strains in each respective outbreak demonstrating that this approach could be used in public health surveillance to include or exclude strains for outbreak detection. However, this is assuming detailed and standardised methods to account for the systematic difference between the two forms of sequencing data are in place.

4.4.2 Outbreak associated with contaminated raw pet food.

The outbreak strain of STEC O157:H7 in this study and the strain Sakai reference strain ^[75] belonging to two different sub-lineages, sub-lineages Ic and Ia, respectively and were isolated 20 years apart in geographically distinct regions, UK and Japan respectively. The prophage commonality shared between the strain Sakai and the outbreak strains indicates stability of some the non-*stx*-encoding prophage content over time and space (Figure 4.3). However, in contrast, the variation of *stx* profiles and the mobility of *stx*-encoding phage to strains belonging to the same lineage that are globally distributed but also in closely related strains at the local level is well documented ^[80,103]. Previous studies charting the evolutionary history of STEC O157:H7 propose the progenitor strain of lineage I harboured the *stx2c*-encoding prophage only ^[80,81]. During its evolutionary history, the strain Sakai appears to have lost the *stx2c*-encoding prophage and acquired a *stx1a*-encoding prophage (which is genetically similar to the *stx*-negative prophage 11 in this study) and a *stx2a*-encoding prophage, although the order of these events cannot be determined within this project. It is well described that sub-lineage Ic harbouring the *stx2a*-encoding prophage emerged in the UK in the 1990s resulting in the change from PT from PT32 to PT21/28 ^[80,81,103]. However, in this instance, the *stx2a*-encoding prophage (prophage 12) and Sp5 (Sakai's *stx2a*-encoding prophage) share only 40% of hashes (transformed nucleotide sequences) via mash and both have different SBI sites (*argW* and *wrbA*, respectively).

Acquisition and loss of *stx*-encoding bacteriophages and non-*stx*-encoding bacteriophages clearly contributes to the evolutionary history of STEC ^[81,203,204],

possibly in response to selection pressures within the ruminant reservoir. A better understanding of the phenotypic expression of the microevolutionary events in the prophage component of the genome may reveal interesting insights into how new strain emerge and persist in the animal reservoir and human population.

4.4.3 Outbreak associated with consumption of raw drinking milk.

Within this outbreak, pairwise comparisons also revealed several common non-*stx*-encoding prophages found in strains of STEC O157:H7 that are temporally and geographically distinct ^[170]. This finding provides further evidence that certain prophages are stable within the STEC O157:H7 chromosome and perhaps can no longer be induced. However, in contrast to the above, some STEC prophages which have previously been described as non-inducible have been shown to be mobile ^[75,111]. Comparing the outbreak strains to other strains from the same sub-lineage and to strains more closely related in time and space, shared more prophages in common than those strains that were phylogenetically, geographically and temporally distinct ^[170,196]. Loss and acquisition of prophage content may be influenced by both time and geographical setting.

Previous studies have described within sub-lineage IIb the acquisition of *stx2a*-encoding prophages after the loss of *stx2c*-encoding prophage exhibiting similar sequences to *stx2c*-encoding bacteriophage at the same SBI ^[103,170]. However, this is the first report of this phenomenon occurring in sub-lineage Ic. Strains of STEC O157:H7 harbouring more than one *stx2a* prophage have been described previously ^[170], but again this is the first report of the acquisition of two different *stx2a*-encoding prophages in this sub-lineage.

Strains harbouring *stx2a* only belonging to sub-lineage Ic are significantly more likely to be associated with severe clinical outcomes than those strains harbouring *stx2a* only in sub-lineage IIb [205]. However, those previous records only account for a single copy of *stx2a*-encoding prophage; the presence of two different *stx2a*-encoding prophages on the same chromosome could play a role in the high pathogenicity observed from cases within sub-lineage Ic. Still, there are several other important factors to consider that might drive this observed pathogenicity including the *stx*-encoding prophage SBI and factors such as the gene content of those prophages.

4.4.4 Outbreak associated with contaminated mud-based obstacle course.

Two of the samples within this outbreak showed a noticeable difference to their Illumina counterpart (588889 and 591229). Analyses suggested the loss of a small 8kbp Insertion element from the chromosome in these two samples. Although the IS element was small, this impacted the alignments of the reads of these samples to the Sakai reference genome. As the reference genome contained this paralogous IS element and the aforementioned samples did not, it led to reads from other paralogous sequences incorrectly aligning to the element in question causing an artificial increase in the number of undetermined (N) positions during variant calling. This led to an increase in the total number of variants between the two samples and all others in this outbreak between the Illumina and Nanopore equivalent datasets. This raises an important consideration when performing whole genome sequencing on multiple platforms retrospectively. Retrospective sequencing on a different

platform compared to the original platform could lead to a disparity in genome features observed. Sampling, isolation and culture of the strains should occur in a short timeframe to reduce the bias introduced from storage and any subsequent changes post culture.

Comparing the *stx2a*-encoding prophages from the outbreak strains to a *stx2a*-encoding prophage of another known PT8 STEC O157:H7 from sub-lineage IIb (315176) showed very little variation despite a difference in geography and a four-year difference in isolation time and (Figure 4.13). Both cases fall within sub-lineage IIb, the *stx2a*-encoding prophage is located with the SBI *sbcB* which is uncommon for *stx2a*-encoding prophages when compared to the wider STEC O157:H7 population [81,116].

Sub-lineage IIb is associated with exposure to sheep [190]. The temporally related samples in the outbreak are also part of sub-lineage IIb and associated with sheep [103] as before the outbreak it was known the location of the race was home to a flock of sheep [191]. Despite their temporal and geographical differences between the above sample (315176), the lack of observed variation between the *stx2a*-encoding prophages could suggest that *stx*-encoding prophages maybe indicative of a particular host reservoir in this case in the gastrointestinal tract of sheep however, more evidence is required.

Comparing the total prophage content revealed structural differences within the prophage content across several samples including a chromosomal rearrangement affecting the relative structure of several prophages (Figure 4.10).

A pairwise comparison has revealed that in STEC O157:H7 strains temporally and geographically distinct share several non-*stx*-encoding prophages [103,170] agreeing

with previous outbreaks that certain prophages are stable within the STEC O157:H7 genome and perhaps can no longer be induced. However, in contrast to the above, some prophages which have previously been described as non-inducible have been shown to be mobile [75,111].

The majority of the microevolutionary events observed occurred within the prophage content of the genome. This level of variation within outbreak strains demonstrates the plasticity of the STEC O157:H7 genome however, without a benchmark or quantitation of genome variation of more outbreaks it is difficult to predict the expected amount of variation of prophage content within STEC populations.

Prophage variation and accessory genome variation of STEC O157:H7 could be further driven by the type of reservoir and transmission route to the human host such as environmental transmission rather than foodborne or person to person transmission.

4.4.5 Outbreak associated with animal contact during a lambing event.

The strains associated with the outbreak linked to the lambing event [61] showed more variation in terms of large structural rearrangements compared to other outbreaks described in this chapter (Figure 4.16), not only were inversions observed but also an inverted translocation. These types of events have been described previously in *Salmonella* species [206] and STEC O157:H7 [207]. However, there is a question to whether these events are genuine or an artifact. One possible explanation is due to the long-term storage of these strains. The strains of this outbreak were sequenced on Nanopore platforms after being stored within Dorset agar for six years. It has been suggested that large structural rearrangements can

occur due to increased stress to the cell following long-term storage of strains. However, despite this, the type (inversion/inverted translocation), location and breakpoints of each rearrangement are consistent with and have been observed in other outbreaks described in this chapter and in other studies [207]. Comparing the prophage content within this outbreak again showed more variation than previously described outbreaks this is due to the recombination within the chromosome described above. As each of the endpoints are within prophage regions (Figure 4.16) and via homologous recombination separate modules of prophages are effectively exchanged creating the observed variation (Figure 4.17).

4.5. Chapter 4 conclusions.

Comparing Illumina with Nanopore sequencing data highlighted the difficulty of SNP typing associated with attempting to integrate both technologies. However, the analysis of the Nanopore data confirmed the close genetic relatedness demonstrated by the Illumina data. It also highlights the need for correct masking of mobile genetic elements and methylated sequences relative to the reference genome used for variant calling. Despite all samples within each respective outbreak being within a 5-SNP single linkage cluster using Illumina sequencing data, analysis of the assembly-based Nanopore sequencing data revealed more variation in both plasmid and prophage content of the genomes than previously expected. With diversity within an outbreak being at the nucleotide level (SNPs) through to synteny level where large chromosomal rearrangements have been observed. With the development of long-read sequencing technology, we can now detect and describe both large- and small-scale structural variation and explore the effect on

phenotype. The ability to characterise the accessory genome in this format is the first step to understanding the significance of these micro-evolutionary events and their impact on the evolutionary history, virulence, and potentially the likely source and transmission of this zoonotic, foodborne pathogen.

Chapter 5:

The use of long-read sequencing to detect and characterise large chromosomal rearrangements in STEC O157:H7 genomes.

5.1. Chapter 5 background.

Genome variation occurs in many forms, including nucleotide variation, small-scale gene variation but also large-scale variation [208]. These large-scale events have been termed as large chromosomal rearrangements (LCRs) and have been observed occurring in bacterial chromosomes [207].

LCRs were considered to occur based on evidence from pulse-field gel electrophoresis (PFGE) experiments [209,210,211]. However, the third-generation sequencing era (long-read sequencing) has finally demonstrated that bacterial genomes are much more plastic than previously expected, ranging across multiple species, including radiation resistant bacteria [212], *Bordetella pertussis* [213]. More specifically LCRs have been shown to occur in gastrointestinal pathogens including *Shigella* species [214], *Salmonella* species [206], *E. coli* [125,126] and STEC [207].

It has been demonstrated that LCRs within STEC tend to occur between two homologous prophages and multiple pairings are common in STEC genomes [207].

There is much less information on the occurrence and frequency of LCRs within the short temporal and geographical scale such as within outbreaks.

Generating complete and closed assemblies is difficult using short-read sequencing technologies because reads shorter than the length of homologues repeats within the STEC genome which prevents complete contigs being produce and thus limits the ability to detect and characterise LCRs. The aim of this chapter is to utilise long-read sequencing to detect and characterise LCRs within outbreaks of STEC O157:H7.

5.2. Chapter 5 methods.

5.2.1 Strain selection and culture.

Strains sequenced within this chapter are the same as those in section 4.2.1.

All strains selected were cultured from long-term storage at UK Health Security Agency (UKHSA) [known previously as Public Health England (PHE)] using 4ml of double-strength nutrient broth (Oxoid) and incubated at 37°C overnight.

5.2.2 Illumina-based DNA extraction, library preparation, sequencing and data processing.

Genomic DNA was extracted from cultures of STEC O157:H7 using the QIAAsymphony system (Qiagen, Hilden, Germany). The sequencing library was prepared using the Nextera XP kit (Illumina, San Diego, USA) for sequencing on the HiSeq 2500 and NextSeq 1000 platforms (Illumina, San Diego, USA), each run with the fast protocol. FASTQ reads were processed using Trimmomatic v0.27 ^[140] as previously described ^[195] to remove sequencing barcode/index adapters. Also, bases from reads were removed from the trailing end that fell below a PHRED score of 30. If the read length post-trimming was less than 50bp, the read and its pair were discarded.

5.2.3 Nanopore-based DNA extraction, library preparation, sequencing and data processing.

High-molecular weight (HMW) genomic DNA was extracted and purified using the Revolugen Fire Monkey HMW DNA extraction kit (RevoluGen, UK) to manufacturer's

instructions. DNA QC and quantification was performed using a Qubit and the HS (high sensitivity) dsDNA assay kit (ThermoFisher Scientific, Waltham, USA) to manufacturer's instructions. Library preparation was performed using the rapid barcoding kit (SQK-RBK004) with each extract being barcoded/indexed via transposase-based tagmentation. Sequencing was performed on the MinION Mk1B for 48 hours as previously described ^[170,196].

Raw FAST5 data was produced from the sequencer before basecalling and demultiplexed using Guppy (versions v3.2.4 through to v3.2.10 all using the FAST model) (Oxford Nanopore Technologies, Oxford, UK) into FASTQ format. Demultiplexing was re-performed using Deepbinner v0.2.0 ^[182]. Sequencing run and QC metrics were generated using Nanoplot v1.8.1 ^[141] and read trimming was performed using Porechop v0.2.4 ^[142] to remove the Nanopore y-adapter and each barcode index. Finally, read filtering using Filtlong v0.2.0 ^[143] with the following parameters, min length=1000 bp, keep per cent=90 and target bases=275 Mbp, to generate approximately 50x coverage of the STEC genome with the longest and highest-quality reads.

5.2.4 *De novo* assembly, correction, reorientation, and annotation.

The filtered Nanopore FASTQ file with the 50x coverage of longest reads were assembled using Flye v2.8 and ^[144] with the minimum overlap length (-m) set to 10000 and the --meta component enabled. Correction (polishing) of the assemblies was performed in a three-step process. Firstly, using Nanopolish v0.11.3 ^[145] with the parameters set to --methylation-aware=dcn,dam, --min- candidate-depth=10 and --min-candidate-frequency=0.1 using Nanopore reads and a self-alignment generated

from Minimap2 v2.7 ^[146] and Samtools v1.7 ^[147], secondly, using Pilon v1.22 ^[148] with parameters set to --minmq=0, --minqual=0 and --mindepth=0.05 with Illumina reads and finally Racon v1.3.3 ^[150] with parameters set to --error-threshold=0.3 and --quality-threshold=10 again using Illumina reads. Nanopore based alignments were generated using Minimap2 v2.17 ^[146] and Samtools v1.7 ^[147]. Illumina based alignments using BWA-MEM v0.7.17 ^[149] and Samtools v1.7 ^[147].

If each assembly contig was circularised and closed, they were re-orientated to start at the *dnaA* gene (GenBank accession no. NC_000913) from *E. coli* K-12, using the --fixstart parameter in Circlator v1.5.5 ^[184]. Prokka v1.13 ^[155] was used to annotate the final assemblies.

5.2.5 Prophage detection, excision and comparison.

Prophage sequences were detected and extracted from each samples' chromosome as described in Shaaban *et al* 2016 ^[81] and Yara *et al* 2020 ^[170]. Firstly, Propi v0.0.1 ^[81] was used to detect and extract prophages from each sample's corrected assembly. Assemblies were then manually curated for prophage detection and excision followed by prophage annotation using Prokka v1.13 ^[155]. Finally, Clinker v.0.0.27 ^[215] was used to generate pairwise alignments for visualisation.

5.2.6 Detection and characterisation of chromosomal structural variation.

Chromosome synteny was compared by aligning chromosomes using Circos v0.69-8 ^[158]. Once samples in one chromosome were aligned, structural differences could be determined and further characterised using Artemis v18.1.0 ^[197].

To determine if there were multiple isoforms within each sample's Nanopore reads (FASTQ). The Nanopore FASTQ for a given outbreak's sample in one isoform was aligned to a finalised assembly with a different isoform using Minimap2 v2.17 ^[146] and Samtools v0.7.17 ^[147] and vice versa. Using Tablet v1.17.08.17 ^[166], the alignments were visualised and the breakpoints identified in each isoform. Once breakpoints were identified relative to each isoform, those positions were used with Samtools v0.7.17 ^[147] view to isolate reads that only align across both ends of each prophage breakpoint. Any reads that did align across a given set of breakpoints must share the same size as it exists in the FASTQ file and not clipped to be considered. Homologous regions were detected using the Nucmer feature of mummer v4.0.0 ^[216] by aligning a given selected assembly to itself where alignment lengths where $\geq$ 2.5kbp.

5.3. Chapter 5 results.

5.3.1 Chromosomal inversions.

In total there were four individual inversions detected relative to other samples within each respective outbreak cluster. The inversions ranged from 0.309Mbp to 1.468Mbp in size. All breakpoints of each inversion were associated with prophage regions of the STEC O157:H7 genome. Each individual inversion observed is detailed below:

Sample 432301, associated with raw drinking milk outbreak ^[59]:

There was a 0.65 Mbp inversion between prophage 5 (located at *potC*) and prophage 7 (located at *yebW*) (Figure 5.1) relative to other outbreak samples. At either edge of the inversion, prophages 5 (*potC*) and 7 (*yebW*) both share a 10.05kbp homologous region (97.7% sequence similarity). The 10.05kbp region contained the structural bacteriophage encoding genes such as host specificity proteins, tail genes, and several hypothetical genes.

When comparing each sample's read sets to the opposite isoform assembly e.g. sample 438729 reads aligned to sample 432301's assembly, this particular isoform appears to disagree with approximately 25% (63/258) of reads. This translates to approximately 75% (195/258 reads) of reads agree with the assembled 432301 isoform. This demonstrates that reads that can align to both isoforms exist with a single read set.

438729

432301

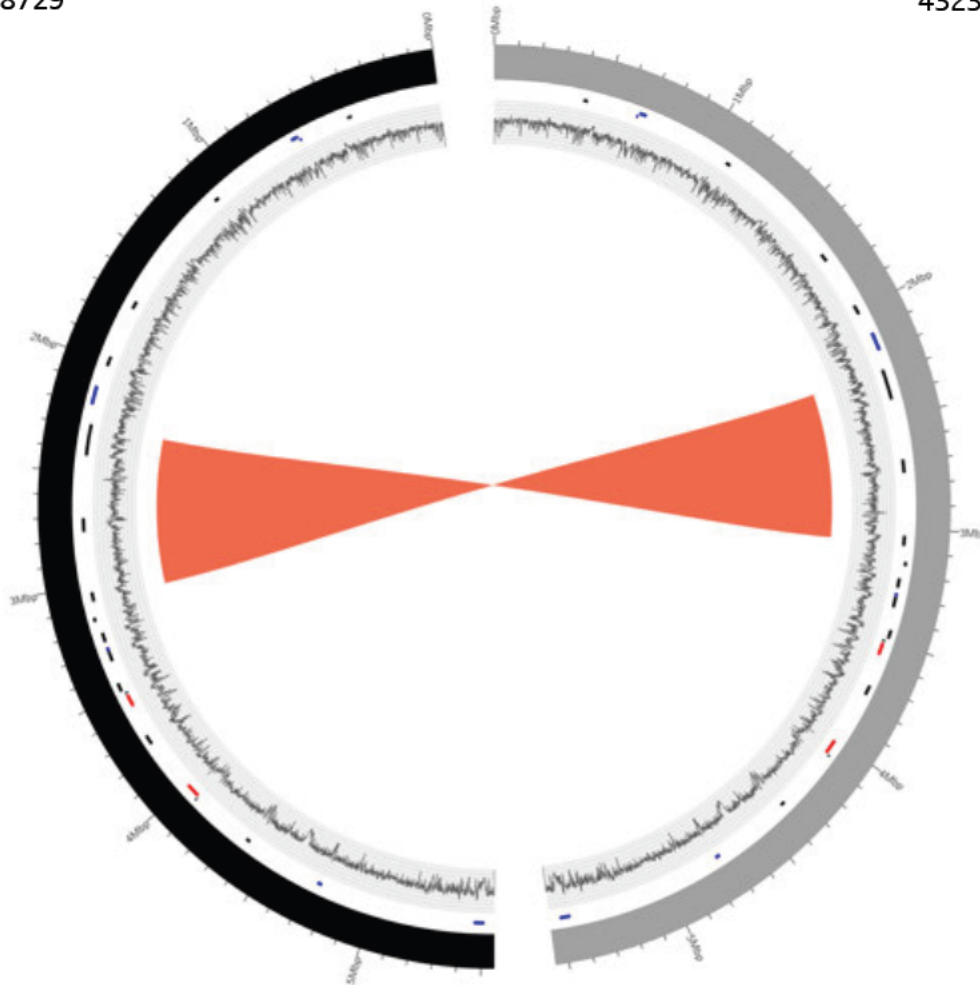


Figure 5.1 – Circos plot showing the 0.65Mbp chromosomal inversion (red link) between samples 432301 (grey hemisphere) and 438729 (black hemisphere). Also detailing prophage regions as black blocks, prophage-like regions as blue blocks and *stx*-encoding prophages as red blocks. Inner ring shows GC content at a 1kbp sliding window.

Samples 311778 and 401794, associated with an environmental contamination of a mud-based obstacle course ^[191]:

Both genomes of 401794 and 391778 showed a two inversion events, one nested within the other, relative to the other strains within this outbreak. The wider inversion, approximately 1.468Mbp in size, occurred with the breakpoints within prophage 4 (*yccA*) and prophage 12 (tRNA-Ser) (Figure 5.2). A 12.6kbp homologous sequence spanning host specificity proteins, terminase gene and structural genes such as head, tail and portal encoding genes was the homologous (99.55% sequence similarity) sequence identified at prophage 4 (*yccA*) and prophage 12 (tRNA-Ser), the breakpoints of the 1.468Mbp inversion.

Taking two exemplar samples (one chromosome in each different chromosomal isoform), 401794 and 588888. The wider inversion at prophage 4 (*yccA*) and prophage 12 (tRNA-Ser) the associated breakpoints, approximately 40% of reads (27/69) from the read set from 401794 aligned to the opposing isoform in the assembly of 588888. Also vice versa 30% of reads (78/260) from sample 588888 aligned to the opposing isoform in the assembly of 401794. This shows that both isoforms exist within the read sets of both exemplar samples.

The nested inversion, approximately 0.452Mbp in size, occurred between prophage 9 (*ompW*) and prophage 10 (*rspR*) (Figure 5.2). Each prophage shared a 20kbp homologous sequence (99.57%) which again included prophage structural genes such as head, tail and portal encoding genes.

The nested inversion at *rspA* and *ompW* prophages, approximately 58% of reads (214/507) from the 588888 read set agreed with the opposing isoform in the assembly of 401794 demonstrating that this isoform exists in both exemplar samples.

588888

401794

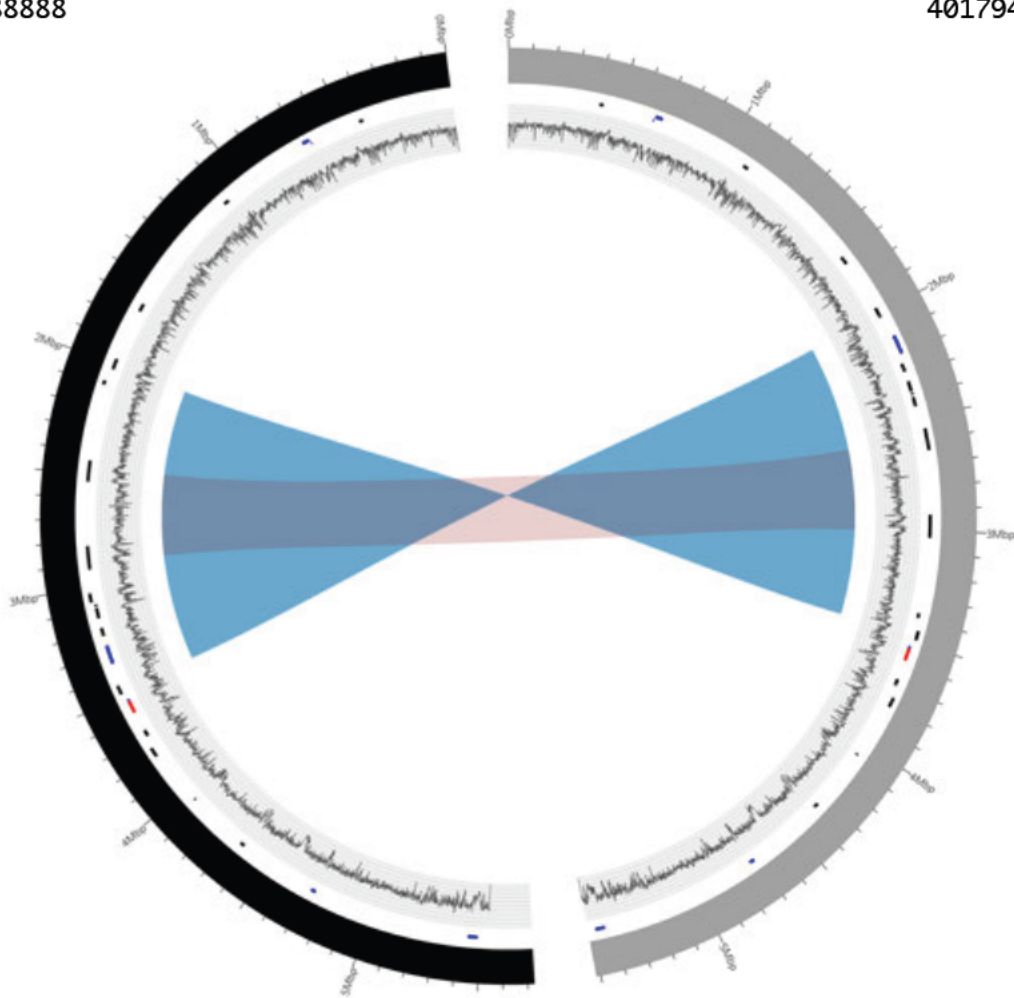


Figure 5.2 – Circos plot showing both 1.46 Mbp (blue link) and 0.45 Mbp (red link) chromosomal inversions between samples 401794 (grey hemisphere) and 588888 (black hemisphere). Also detailing prophage regions as black blocks, prophage-like regions as blue blocks and *stx*-encoding prophages as red blocks. Inner ring shows GC content at a 1kbp sliding window.

Sample 60202, associated with direct animal contact ^[61]:

Comparing the chromosome of sample 60202 to the chromosome 60189, there was a large 309kbp chromosomal inversion between the prophages located at prophage 6 (*rspA*) and prophage 9 (*ompW*) which also spans the terminus of the chromosome (Figure 5.3). At either edge of this inversion, both prophages share a 11kbp (99.24%) homologous sequence spanning across the structural genes of the prophage head, tail, portal and host specific encoding genes. Approximately 31% of the 60202 (64/204) read set from 60202 agreed with the opposing isoform in the assembly of 60189, again demonstrating that both isoforms exist in both exemplar samples.

5.3.3 Chromosomal inverted translocation.

Another recombination event observed was a single 248.6kbp inverted translocation in sample 60202. As the same as the inversions detailed previously, the breakpoints were associated with the prophages of the STEC O157:H7 genome. 60202 has a 248.6kbp inverted translocation of the chromosome involving the translocation of two prophages (partial prophage located at *ompW* and full prophage located at *roxA*) and one prophage-like element (the element that is SPL2 in the Sakai O157:H7 reference genome) (Figure 5.3). The breakpoints of the inverted translocation was formed between the prophage located at *potC* and the *stx2c*-encoding prophage (located at a transposase sequential to *sbcB*). Approximately 39% of the 60202 read set agreed with the opposing isoform in the assembly of 60192 where the translocation has not occurred. This also demonstrates that both isoforms exist in both exemplar samples.

60198

60202

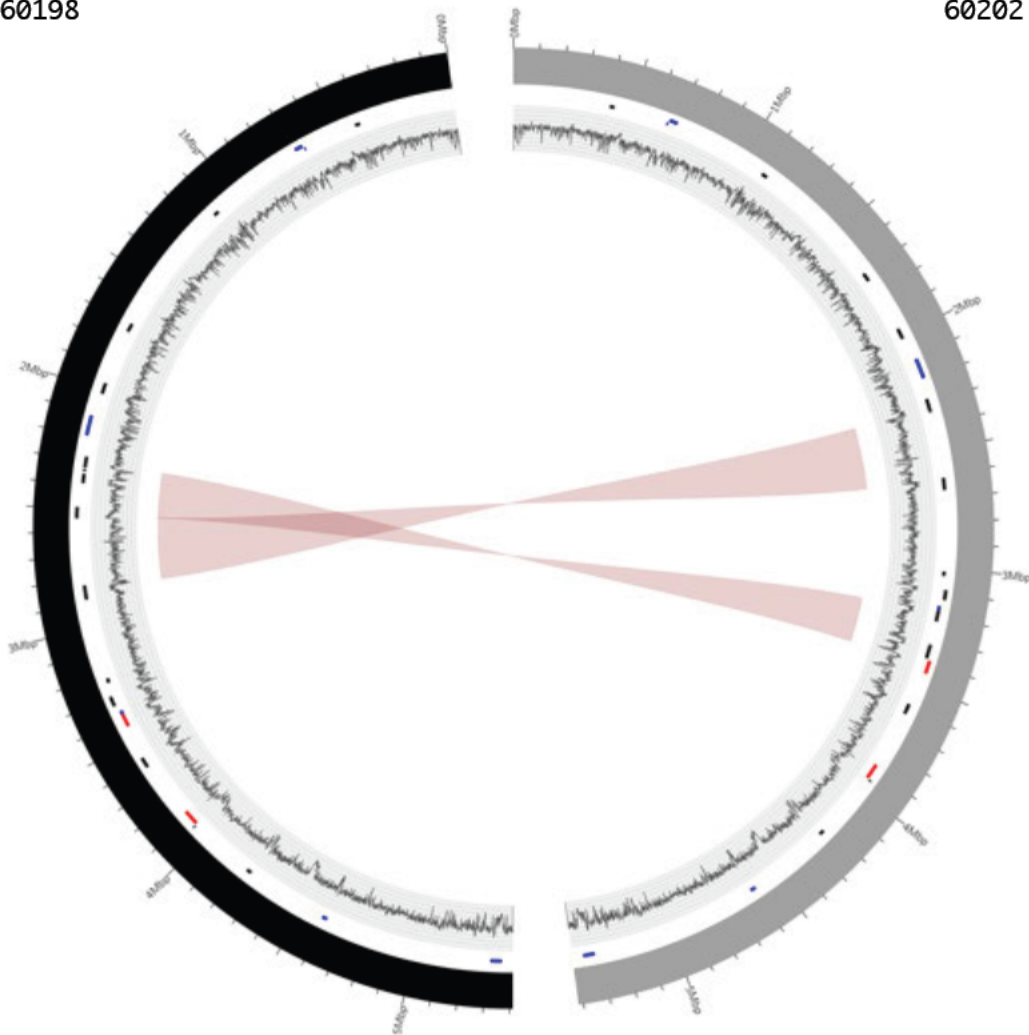


Figure 5.3 – Circos plot showing both 0.309Mbp inversion (red link) and 0.248Mbp inverted translocation (red link) between samples 60202 (grey hemisphere) and 60198 (black hemisphere). Also detailing prophage regions as black blocks, prophage-like regions as blue blocks and *stx*-encoding prophages as red blocks. Inner ring shows GC content at a 1kbp sliding window.

5.3.3 Homologous prophage regions.

Taking genomes where large chromosomal rearrangements have occurred, these genomes were scanned for homologous regions ($\geq 90\%$ sequence similarity) and $\geq 2500\text{kbp}$ in length to search for locations in which homologous recombination could possibly occur.

Sample 432301 associated with the 0.65 Mbp chromosomal inversion harboured 101 such regions totalling 273,557 bp, ranging from 2,549 bp to 16,881 bp in size (average 5,417 bp) and 90.74% to 99.78% (average 96.9%) sequence similarity (Figure 5.4). Of the regions described, 82 were located in prophage regions (81%) totalling 232,215 bp, ranging from 2,549 bp to 16,881 bp in size and 90.74% to 99.78% (average 96.38%) sequence similarity (Figure 5.4)

Sample 401794 associated with nested inversion harboured 81 of these regions totalling 206,096 bp, ranging from 2,546 bp to 17,825 bp (average 5,088 bp) in size and 90.56% to 99.58% (average 96.9%) sequence similarity (Figure 5.5). Of the regions described, 58 were located in prophage regions (71%) totalling 158,206 bp, ranging from 2,546 bp to 17,825 bp (average 5,455 bp) in size and 90.56% to 99.58% (average 96.2%) sequence similarity (Figure 5.5).

Sample 60202 associated with both an inversion and inverted translocation harboured had 55 of these regions totalling 137,248 bp, ranging from 2,549 bp to 14,613 bp (average 4,990 bp) in size and 94.17% to 99.63% (average 97.6%) sequence similarity (Figure 5.6). Of the regions described, 34 were located in prophage regions (61%) totalling 92,856 bp, ranging from 2,549 bp to 14,613 bp in size and 94.17% to 99.45% (average 96.9%) sequence similarity (Figure 5.6)

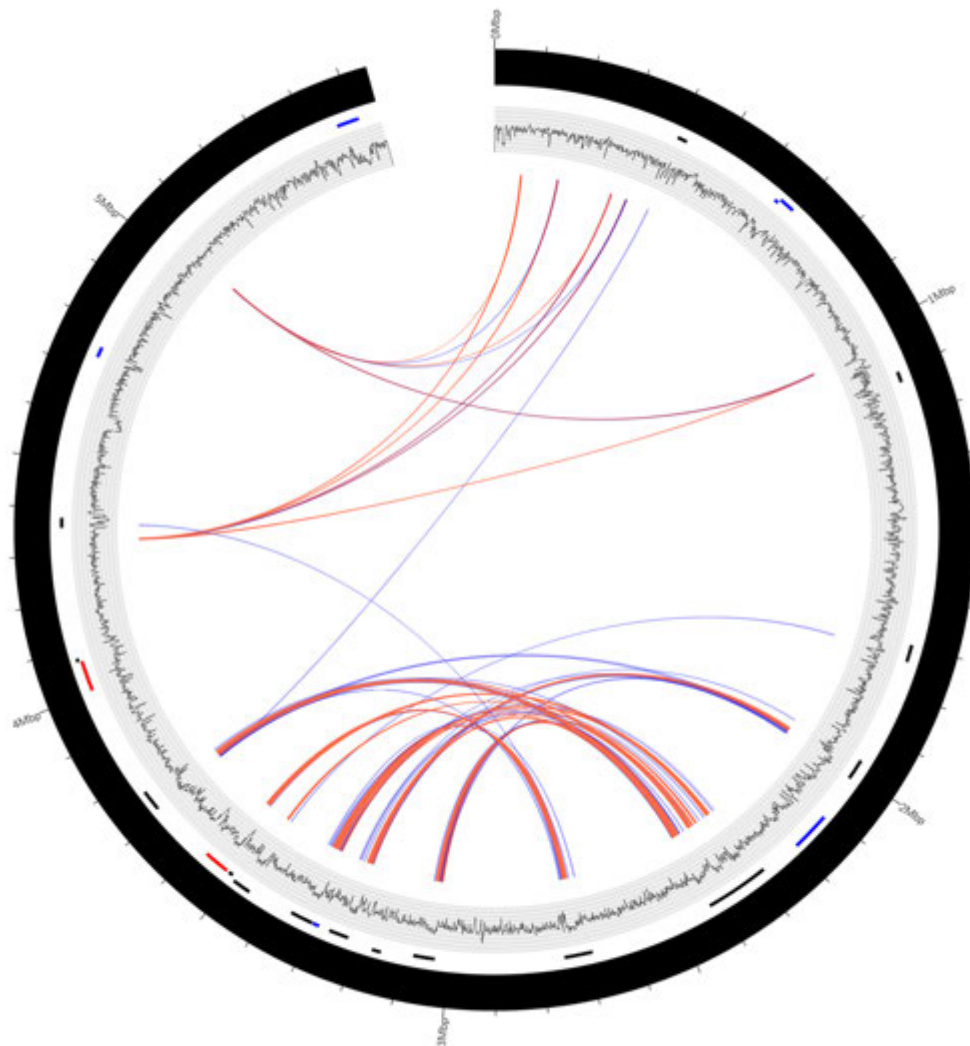


Figure 5.4 – Circos plot showing the chromosome of sample 432301 (black ring), highlighting regions of the genome (links) that are $\geq 90\%$ sequence similarity and $< 5\text{kbp}$ in size (blue) or $\geq 5\text{kbp}$ in size (red). Also detailing prophage regions as black blocks, prophage-like regions as blue blocks and *stx*-encoding prophages as red blocks.

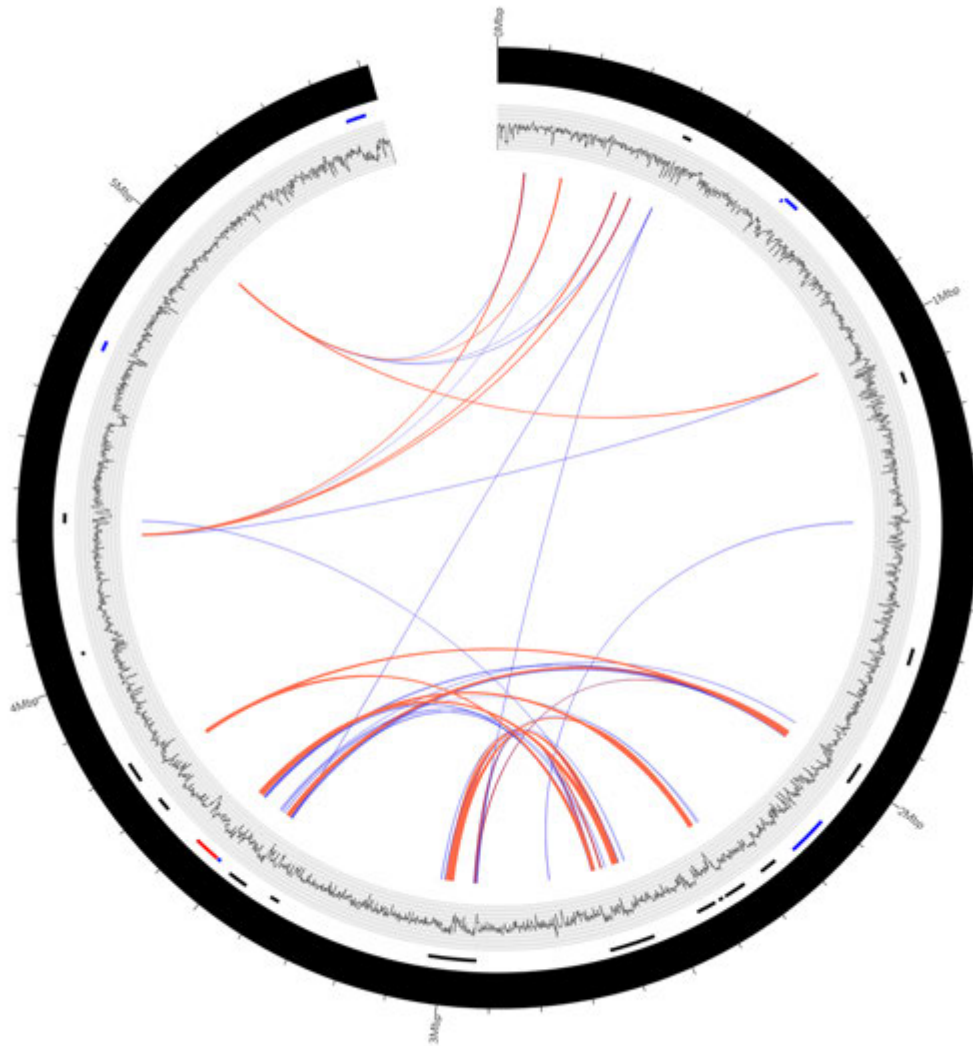


Figure 5.5 – Circos plot showing the chromosome of sample 401794 (black ring), highlighting regions of the genome (links) that are $\geq 90\%$ sequence similarity and $< 5\text{ kbp}$ in size (blue) or $\geq 5\text{ kbp}$ in size (red). Also detailing prophage regions as black blocks, prophage-like regions as blue blocks and *stx*-encoding prophages as red blocks.

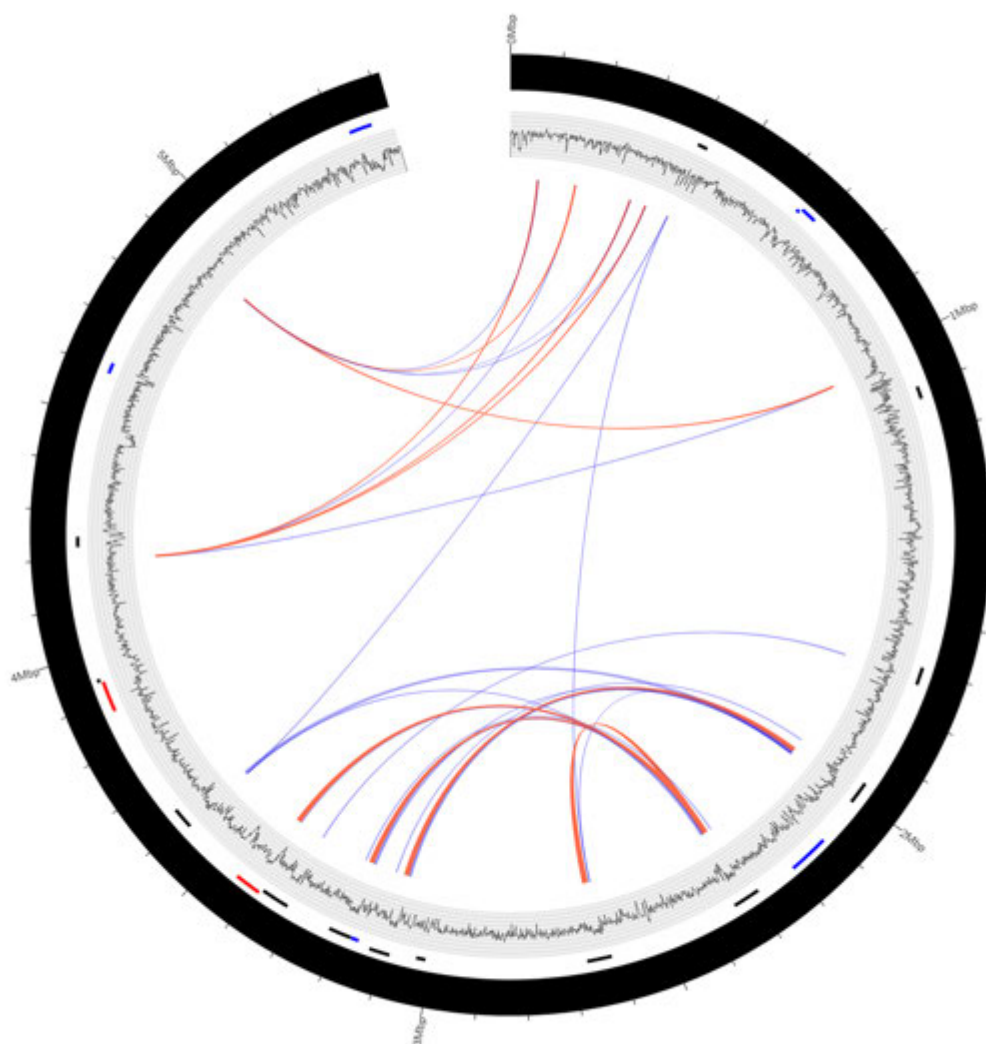


Figure 5.6 – Circos plot showing the chromosome of sample 60202 (black ring), highlighting regions of the genome (links) that are $\geq 90\%$ sequence similarity and $< 5\text{kbp}$ in size (blue) or $\geq 5\text{kbp}$ in size (red). Also detailing prophage regions as black blocks, prophage-like regions as blue blocks and *stx*-encoding prophages as red blocks.

5.3.4 Compound prophages.

Examining prophage content in each of the outbreak samples characterised revealed the presence of either prophages in tandem with other prophages with no discernible chromosomal sequence between them or prophages integrated into other prophages, both termed “compound” prophages onwards. The compound prophages observed ranged from approximately 60kbp to 145kbp in size and ranged from two to four separate prophages per compound prophage (Table 5.1). Samples from the outbreaks associated to raw pet food and raw drinking milk harboured the largest compound prophages averaging 145,057bp and 142,528bp respectively. Both outbreak samples associated with the environmental exposure and direct animal contact harboured two compound prophages at averages in size of 96,362bp, 105,848bp and 60,650 and 91,109kbp respectively (Table 5.1).

Outbreak association	Number of compound prophages	Average length of compound prophage (bp)	Contains approximately n unique phage sequences	Integration site
Raw pet food	1	145,057	4	<i>ompW</i>
Raw drinking milk	1	142,528	4	<i>potC</i>
Environmental exposure	2	96,362 and 105,848	2 and 3	<i>rspR</i> and <i>ompW</i>
Direct animal contact	2	60,650 and 91,109	2 and 3	<i>ompW</i> and <i>rspR</i>

Table 5.1 – Table detailing compound prophages observed within outbreaks sequenced in this thesis. Includes number of compound prophages per strain, average size across respective outbreak samples, associated integration sites and an approximation of how many unique phage modules per compound phage.

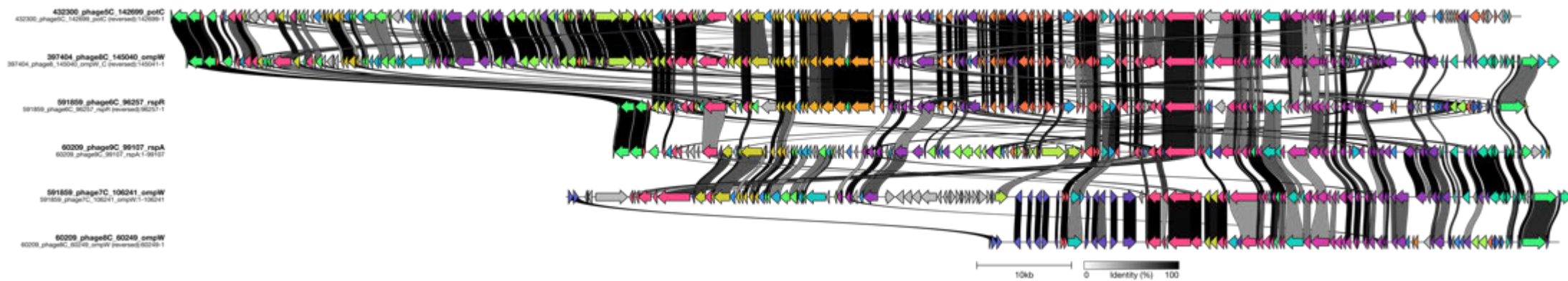


Figure 5.7 – Pairwise clinker alignment of exemplar compound prophages detected in outbreak samples. Labels detail sample ID, phage number within that sample, prophage size (bp) and integration site. Black links denote genes of high sequence similarity, grey links denote partial similarity.

5.4. Chapter 5 discussion.

5.4.1 Prophage boundaries of LCRs.

All the observed large chromosomal rearrangements (LCRs) shared prophage sequences at their boundaries (Figures 5.1, 5.2 and 5.3). As most of these prophages are Lambdoid in terms of structure and gene content ^[111,120,216] they share genes encoding various bacteriophage structures (head, tail fibres, sheath etc). This combined with the high number of prophages harboured by STEC O157:H7 ^[75,207] results in many large homologous repeats throughout the STEC O157:H7 chromosome with homologous regions >2.5kbp occurring up to 82 times within prophage regions in an exemplar single chromosome (Figures 5.4, 5.5 and 5.6). In all LCRs observed, the rearrangements took place over the terminus of the chromosome with one prophage breakpoint being on either side of the terminus (Figures 5.1, 5.2 and 5.3). In the samples sequenced, approximately 66% of the STEC O157:H7 prophage content resides within 1Mbp either side of the chromosome terminus making this region of the chromosome very prophage rich. As a result, the terminus region contains many large prophage homologous sequences and most likely will increase the opportunity for LCRs to occur.

There have been two previous reports of LCRs occurring in STEC, both in relation to colonisation of cattle ^[124,207]. One describes five LCRs of 250kbp to 1.4Mbp occurring across the terminus with homologous prophage sequences providing the breakpoints detected by PFGE profiles. The second observing LCRs in both inversions and a duplication detected with similar WGS methodology and optical mapping.

By their nature, for LCRs to occur they would require DNA to be split at their phage boundaries (double-strand breaks). Repair of these breaks are most likely Rec-mediated (*RecABCD*) as most prophages ^[219], including *stx*-encoding prophages, in the STEC O157:H7 chromosome harbour similar recombinases ^[120]. The SOS response in STEC drives the expression of Shiga-toxin from *stx*-encoding prophages, similarly the SOS response triggers recombinational-based DNA repair mechanisms in bacteria. Therefore, the theoretical frequency of LCR occurrence and Shiga-toxin expression could be proportional ^[207]. It is also undetermined if the repair mechanism is limited to *RecBCD* or host repair or if prophages drive their own DNA-based repair mechanisms to support the formation of LCRs. Secondly, what effect of stresses leading to double strand breaks have on the proportion and frequency of the LCRs? It is more likely that an increase in double-strand breaks from stresses will increase the SOS response and drive *RecBCD* -based repair of the breaks leading to more LCRs.

Typically, LCRs within the STEC O157:H7 chromosome span the terminus region. It is possible that LCRs can occur outside this region. However, this likely generates an imbalance in DNA replication where one replication strand has a longer journey to the terminus, and one has a shorter journey. This imbalance may impact the bacterium's ability to survive in the animal reservoir or the environment. We might be only observing LCRs across the terminus because this does not adversely affect DNA replication as much compared to LCRs that do not cross the terminus equally. The effect of replicore imbalance of the *Salmonella* Typhi chromosome due to recombination has been investigated ^[206]. Extreme imbalance significantly affected gene expression and reduced growth rates of the bacterium. This phenomenon has

also been observed in *E. coli* but not specifically STEC [218], where extreme replicore imbalance significantly reduced growth times of the bacterium.

Fitzgerald *et al* 2021^[207] also detected an inverted duplication affecting a *stx*-encoding prophage and demonstrated that the observed 220kbp inverted duplication appeared to significantly increase the expression and production of Stx and a reduction in Type III secretion system expression. This generates several questions, surrounding the type and location of inversions, duplications and translocations and their effect on expression of virulence factors in STEC O157:H7 and expression of *stx* both in the colonisation of the animal host within the reservoir and colonisation in infectious state in humans. There is no evidence as yet that LCRs provide a fitness advantage in a for the *E. coli* bacterium.

Finally, LCRs most likely drive genetic diversification of prophages that are involved with modules from prophage exchanged with another at the breakpoints of the LCR. The effects of the LCR on the prophages ability to excise and function remains undetermined.

5.4.2 Chromosomal isoforms.

Most public health microbiological high-throughput genomic analyses to date are performed using cost-effective second-generation sequencing technologies producing short-reads [57,111,130,195,220,221]. The chromosomal isoforms described here will remain hidden due to the nature of short-read sequencing resulting in poor contiguity of *de novo* assembly. Going forward detection and determination of relative isoform proportions of LCRs in the STEC genome will need to be performed by long-read sequencing such as the methods described in this thesis.

All the LCRs described in this thesis had a proportion of reads disagreeing with the consensus of the LCR that was assembled. This demonstrates that when LCRs occur, they do not occur in all of the cells within a pure culture. However, given the analyses of three outbreaks in which samples were very closely related via SNP typing, in each case the majority of samples tend to favour a single conformation. Such as 7/10, 22/23 and 10/12 samples for the petting farm outbreak, raw drinking milk outbreak and the outbreak associated with environmental exposure. This suggests that a particular isoform is more stable than another as more native sequencing reads are produced in a particular conformation and the reason why the assembler selected this conformation for finalised contiguous chromosomal structure.

Typically, pure cultures of a bacterium from the same outbreak cluster are thought to be genetically identical bar some possible single nucleotide differences. This analysis has demonstrated that chromosomes of a single sample are dynamic and can undergo chromosomal rearrangements. The implication of this is that within any pure culture of STEC O157:H7, there are a proportion of cells that exist in one isoform of the chromosome and other cells that exist with a different isoform (one that has a relative LCR). Therefore, when considering LCRs it would be prudent to consider a pure culture of cells to not contain the same chromosomal synteny but can support variation.

A possible limitation to these analyses is whether these events occur as frequently *in vivo* as observed *in vitro*. Repeat culture and storage of a strain most likely stresses the strain and could theoretically induce LCR formation. It is possible, therefore, that LCRs are a culture-based artifact, and this hypothesis requires further investigation.

However, this would mean that the events that drive recombination only occur under laboratory conditions, and this would seem unlikely. If anything, stresses *in vivo* can be more extensive and may potentially cause more extreme recombination events that single isolate analysis cannot detect. It is likely that there are a greater number of variants of LCRs within a single culture that are difficult to detect by long-read genomic sequencing due to low copies reads representing those LCRs being present and lack of assembled isoforms to compare to. Revisiting optical mapping could provide the resolution required to observe all types of LCRs within an individual culture. Also, optical mapping can highlight any mis-assembly within long-read-based assemblies. Secondly, creating STEC O157:H7 mutants with decreased ability to repair double strand followed by long-read sequencing could reveal the location and frequency of double strand breaks in the chromosome. Combining optical mapping with a targeted long-read sequencing on DNA repair mutants could allow for the capture of more isoform variants within a pure culture and to understand where the double strand breaks are occurring and if they are random or have specific sites along the chromosome such as in the prophage regions.

5.4.3 Compound prophages.

A small number of prophages observed in the sequenced samples sequenced were much larger than expected (up to 148kbp in size). Further investigation showed instances of a prophage sequential to another prophage with no chromosomal sequence in between or prophages that have integrated inside another prophage. These compound prophages were all found to be close to the terminus of the STEC O157:H7 chromosome and provides multiple points of homology (Figures 5.3, 5.4

and 5.5). The integration site and number of homologous sequences within compound prophages most likely will increase the opportunity for LCRs to occur. Recently, a study detailed an integration sequence (*attB*) within a prophage located at *ompW* harboured by a STEC O145:H28 strain [203]. This could be one mechanism in which compound prophages form, particularly those that exist as a prophage inside another prophage and not sequentially. An important public health consideration from Nakamura *et al* 2021 [203] was whether compound prophage generation could provide a method of accumulating virulence factors, such as Type III secreted effectors (*nle*, *Esp* etc) which are known to be found in prophage sequences [203,222]. In combination with LCRs between compound prophages and other prophages, these micro-evolutionary events could provide a mechanism to accumulate virulence factors and to distribute them across the chromosome, ultimately leading to the emergence of more virulent strains of STEC.

5.4.4 Impact on *de novo* assembly and bioinformatics considerations.

With the consideration that multiple isoforms of a single chromosome can exist in a single pure culture of STEC O157:H7, it raises several bioinformatics related considerations. Firstly, the impact of this on *de novo* assembly, for several samples in this thesis it was noted that assembler could not achieve a contiguous chromosome instead splitting into three contigs each piece predicted to connect at the same point within the chromosome. Therefore, the more equal in proportions that isoforms exist the more difficult it will be for an assembler to complete the chromosome without bioinformatics intervention.

5.5. Chapter 5 conclusions.

The STEC O157:H7 chromosome is much more dynamic and plastic within an outbreak scenario than previously understood with LCRs likely increasing prophage diversity and facilitating the acquisition and distribution of prophage-encoded virulence factors with large chromosomal rearrangements. Further investigations are required to determine the underlying mechanism of LCRs and if they provide any selective advantages that would confer a selective advantage and superior survival strategy for STEC O157:H7.

Chapter 6:

Detection and characterisation of antimicrobial resistance determinants in STEC O157:H7 using long-read sequencing.

6.1. Chapter 6 background.

Clinical management and treatment of patients with infectious disease is confounded by increasing rates of antimicrobial resistance (AMR) [223,224,225,226]. Active surveillance and monitoring of AMR determinants (genes) within bacterial populations is essential to combat the global spread of multidrug resistant (MDR) pathogens. Surveillance of AMR associated with GI pathogens is important as monitoring acquisition of AMR determinants in patients reporting traveller's diarrhoea and, in those reporting, domestic acquisition are vital for predicting patient outcomes and public health action [223,224,225,226].

Surveillance of AMR also allows us to monitor emerging MDR, novel exposures and routes of transmission [227]. Because STEC O157:H7 is zoonotic, surveillance of STEC O157:H7, facilitates monitoring the transmission of AMR determinants from the animal reservoir to humans via the food chain [228,229]. Prior to the implementation of WGS at UK Health Security Agency (UKHSA), AMR detection in STEC was limited, partly because antimicrobial susceptibility testing (AST) was not required as antibiotic treatment of STEC is contraindicated and partly for safety reasons as AST for Hazard Group 3 pathogens carries an increased risk to laboratory staff.

Since 2015, *in silico* AMR profiles of all STEC submitted to UKHSA have been produced in real-time [228,229]. The STEC O157:H7 population can be divided into three main lineages (I, II and I/II) and further subdivided into seven sub-lineages, Ia-Ic, IIa-IIc and I/II [80]. Within the STEC O157:H7 population observed in the UK, domestically associated strains typically belong to sub-lineages Ic and IIb and some in IIa and IIc. Whereas typically travel-associated isolates are mostly associated only with sub-lineages IIa and IIc [98,99].

The aim of this chapter was to utilise long-read sequencing data in combination with short-read sequencing data to determine the occurrence, frequency and genetic context of AMR determinates across the population of clonal complex 11 (CC11) STEC O157:H7.

6.2. Chapter 6 Methods.

6.2.1 Strain selection and culture.

Strains from England of STEC O157:H7 that were submitted to GBRU during the timeframe of 1st January 2016 to 31st December 2020 were routinely processed through the Illumina-based WGS workflows at UKHSA. All strains selected were cultured from long-term storage at UK Health Security Agency (UKHSA) [known previously as Public Health England (PHE)] using 4ml of double-strength nutrient broth (Oxoid) and incubated at 37°C overnight.

6.2.2 Illumina-based, DNA extraction, library preparation and sequencing.

A Qiagen Qiasymphony (Qiagen, Hilden, Germany) was used to extract Genomic DNA from isolates. Library preparation and indexing was performed using the Nextera XT kit (Illumina, Cambridge, UK) before sequencing on an Illumina HiSeq (Illumina, Cambridge, UK) platform to produce 100bp short-read sequence reads.

6.2.3 Illumina-based bioinformatics processing.

Raw FASTQ files (reads) were processed using Trimmomatic v0.27 ^[140] to remove sequencing barcode/index adapters. Also, bases from reads were removed from the trailing end of that fell below a PHRED score of 30. If the read length post-trimming was less than 50, the read and its pair were discarded.

Gene-Finder ^[161], a customised alignment tool which utilises Bowtie2 v.2.3.5 ^[149] and Samtools v.1.8 ^[147] to align reads to a set of AMR gene reference sequences as previously described ^[228,229].

AMR genes were defined as detected based on $\geq 90\%$ nucleotide identity relative to the reference sequence and 100% read coverage, apart from β -lactamase variants that were determined with 100% identity using the reference sequences downloaded from the Lahey (www.lahey.org) and National Center for Biotechnology Information (NCBI) β -lactamase data resources (<https://www.ncbi.nlm.nih.gov/pathogens/beta-lactamase-data-resources>). Known resistance-conferring mutation-based genes and acquired-resistance genes and relevant to β -lactams (including carbapenems), aminoglycosides, chloramphenicol, fluoroquinolones, sulphonamides, fosfomycin, macrolides, tetracyclines, trimethoprim, rifamycins and were included in the analyses. Chromosomal mutations focused on variations in the quinolone resistant determining regions (QRDR)s of *gyrA*, *gyrB*, *parC* and *parE*. If an isolate had resistance to three or more classes of antimicrobial it was defined as multi-drug resistant (MDR).

6.2.4 Phylogenetic tree construction.

The STEC O157:H7 reference genome Sakai (GenBank accession BA000007) was used as a reference and Illumina reads from each sample were aligned using BWA-MEM v.0.7.13 ^[149] and Samtools v.1.8 ^[147]. Variant calling was performed using GATK v.2.6 in unified genotyper mode ^[163]. High-quality SNPs ($>90\%$ consensus, minimum depth $10\times$ MQ ≥ 30) in at least one isolate were extracted for further analysis and imported into SnapperDB v.0.2.8 ^[94]. A whole-genome alignment was

generated from SnapperDB v.0.2.8 ^[94] before processing through Gubbins v.3.2.0 ^[167] to mask any sequences associated with recombination. Finally, a maximum-likelihood phylogenetic tree was constructed by RAxML v.8.1.17 ^[168] and the phylogeny was visualised/annotated using iTol v5 ^[230].

6.2.5 Nanopore-based, DNA extraction, library preparation and sequencing.

The Revolugen Fire Monkey DNA extraction kit (Revolugen, Glossop, UK) was used to extract and purify high-molecular weight genomic DNA. Library preparation was performed using the Rapid Barcoding kit (SQK-RBK004) (Oxford Nanopore Technologies, Oxford, UK). The prepared libraries were loaded onto FLO-MIN106 R9.4.1D flow cells (Oxford Nanopore Technologies, Oxford, UK) and sequenced using the MinION (Oxford Nanopore Technologies, Oxford, UK) for 48 h.

6.2.6 Nanopore-based bioinformatics processing.

Raw sequencing data was produced in a FAST5 format. The raw data was basecalled and de-multiplexed using Guppy v.4.3.4 [FAST model] (Oxford Nanopore Technologies, Oxford, UK) into FASTQ format and grouped in each sample's respective index/barcode. Using Deepbinner v.0.2.0 ^[182], each sample was re-demultiplexed into respective barcodes/indexes. Sequencing run metrics were generated using Nanoplot v.1.8.1 ^[141]. Porechop v.0.2.4 ^[142] was used to remove the barcode and y-adapter from each sample's reads followed by trimming and chimeric reads split. Finally, the trimmed reads were filtered using Filtlong v.0.2.0 ^[143] with the following parameters, min length=1000bp, keep percentage=90 and target bases=

275 Mbp, to generate approximately 50× coverage of the STEC genome (5.5 Mbp) to generate two FASTQ files one for the longest (--length-weight=10) and one for the highest-quality (--mean-q-weight=10) reads.

6.2.7 *de novo* assembly and assembly correction.

A FASTQ file of each sample with 50× coverage of longest reads was assembled using Flye v.2.8 ^[144] with the minimum overlap length (-m) set to 10000 and the --meta component enabled. Assembly polishing (or correction) was performed in a three-step process. Firstly, using Nanopolish v.0.11.3 ^[145] using both the highest-quality filtered FASTQ and the FAST5 files for each respective sample accounting for methylation using the --methylation-aware=dcm and --min-candidate-frequency=0.1. Minimap2 v.2.17 ^[146] and Samtools v.1.7 ^[147] was used to generate the self-alignments for each genome to be processed by Nanopolish v0.11.3 ^[145]. Secondly, Pilon v.1.22 ^[148] was used in conjunction with Illumina FASTQ reads as the query dataset with the use of BWA v.0.7.17 ^[149] and Samtools v.1.7 ^[147]. Finally, Illumina FASTQ reads were used with Racon v.1.3.3 ^[150] also using BWA v.0.7.17 ^[149] and Samtools v.1.7 ^[147] to generate the alignment. Each contig from each assembly was circularised and closed. Using Circlator v.1.5.5 with the --fixstart parameter ^[184] each contig were re-orientated to start at the *dnaA* gene (GenBank accession no. NC_000913) from *E. coli* K-12 and the database of plasmid replicons from PlasmidFinder v.2.1 ^[198]. Finally, all contigs were annotated using Prokka v.1.13 ^[155] with default parameters.

6.2.8 *In silico* plasmid replicon typing, detection and characterisation of AMR determinants.

Each Nanopore-based non-chromosomal contig within the final assembly of each sample underwent plasmid replicon typing using PlasmidFinder v.2.1 ^[198] with the Enterobacteriaceae group selected, minimum identity=90% and minimum coverage = 90% parameters set. The contig which were harbouring AMR determinants were annotated using PGAP build 2022-12-13 ^[156]. Finally, Clinker v.0.0.27 ^[215] was used to generate pairwise alignments for visualisation.

6.3. Chapter 6 results.

6.3.1 General presence of AMR determinates in STEC O157:H7.

Within the study timeframe 14.7% (217/1,473) were found to contain at least one AMR determinant (Table 6.1). Breaking this down by sub-lineage it was noted that there was variability with fewer AMR determinants detected within sub-lineage IIa (9.6%, 31/323), sub-lineage Ia (7.9%, 3/38), sub-lineage IIb (3.3%, 122/440) and sub-lineage Ic (2%, 7/342). In contrast, sub-lineages presenting with the highest proportion of AMR determinants per sample were sub-lineage IIc at 27.7% (122/440), sub-lineage I/II at 35.3% (18/51) and finally sub-lineage Ib at 43.7% (28/64) (Table 6.1). Consequently, these three sub-lineages (I/II, Ib and IIc) were further analysed using Nanopore sequencing to characterise and describe the genomic architecture of the AMR determinants.

6.3.2 Presence of AMR determinants by antibiotic class.

Predicted resistance to β -lactams was associated with 8.8% (129/1,473) of isolates in the study window. This included *bla*_{TEM-1} (n = 116), *bla*_{CTX-M-15} (n = 7), *bla*_{TEM-191} (n = 3) *bla*_{TEM-117} (n = 2) and *bla*_{TEM-190} (n = 1). Three cases reported travel outside the UK before symptom onset (Table 6.1).

Predicted resistance to tetracycline (*tetA*) was associated with 11.3% (166/1,473) of isolates across all sub-lineages within CC11 (Table 6.1).

Sulphonamide resistance was predicted to be in 11.7% (173/1,473) of isolates with 111 isolates containing *sul-1*, 53 containing *sul-2* and 9 with *sul-3* alone (or in combination) (Table 6.1).

Trimethoprim resistance associated with *dfrA* was predicted to be in 7.9% (117/1,473) of isolates with variants including *dfrA-1* (n=66), *dfrA-8* (n=41), *dfrA-12* (n=10), *dfrA-5* (n=1) and *dfrA-14* (n=1) (Table 6.1).

Streptomycin resistance associated with *strA* or *strB* was predicted to be in 11.7% (172/1,473) of isolates of which 70.9% (167/172) were in combination with a set of genes conferring resistance to aminoglycosides. Of the genes conferring resistance to aminoglycosides there were three *aadA* variants observed (*aadA-1*, *aadA-1b* and *aadA-2*) (Table 6.1).

Predicted macrolide resistance was observed in 0.5% (8/1,473) of isolates of which five harboured *ermB* and three harboured *mphA* (Table 6.1).

Quinolone resistance genes either by mutations in *gyrA* or *parC* or acquired resistance genes (*qnrS1*) were observed in 2.9% (43/1,473) of isolates (Table 6.1).

In terms of other AMR determinants 3.4% (51/1,473) of isolates were observed to harbour (alone or in combination) *cml-1*, *catA* or *floR* (Table 6.1).

Lineage	Ia		Ib		Ic		IIa		IIb		IIc		I/II	
	n	%	n	%	n	%	n	%	n	%	n	%	n	%
Total	38	100	64	100	342	100	323	100	215	100	440	100	51	100
Fully susceptible	35	92.1	36	56.3	335	98	292	90.4	208	96.7	318	72.3	33	64.7
B-lactams	0	0	12	18.8	5	1.5	13	4	4	1.9	95	21.5	0	0
Aminoglycosides	3	7.9	27	42.2	0	0	17	5.3	2	0.9	108	24.5	15	29.4
Fluroquinolones	1	2.6	1	1.6	0	0	7	2.2	0	0	32	7.3	2	2
Macrolides	3	7.9	0	0	0	0	1	0.3	0	0	7	1.6	0	0
Trimethoprim	1	2.6	14	21.9	0	0	6	1.9	0	0	96	21.8	0	0
Sulphonamides	2	5.3	27	42.2	5	1.5	17	5.3	0	0	107	24.3	15	29.4
Tetracycline	3	7.9	26	40.6	4	1.2	17	5.3	1	0.5	99	22.5	16	31.4
Chloramphenicol	0	0	0	0	0	0	9	2.9	0	0	48	10.9	1	1
Recent travel outside UK	11	28.9	18	28.1	15	4.4	86	26.6	9	4.2	93	21.1	8	15.7

Table 6.1 - Number of isolates predicted to confer resistance subdivided by CC11 STEC O157:H7 sub-lineage and the number of cases reporting recent travel (within 7 days of onset of symptoms) outside the UK.

6.3.3 Phylogenetic analysis of AMR determinants and analysis of Nanopore-based data in sub-lineage I/II.

Within sub-lineage I/II most of the MDR strains were located within a 50 SNP single-linkage cluster harbouring *strA*, *tetA*, *strB*, *sul-2* (Figure 6.1). These AMR determinants were found to be on a 9kbp Tn10 integron via Nanopore sequencing. This integron was found to be within a prophage located at *clpA* (Figure 6.2).

6.3.4 Phylogenetic analysis of AMR determinants and analysis of Nanopore-based data in sub-lineage Ib.

The majority of the isolates within sub-lineage Ib contained one of two AMR profiles, either *bla*_{TEM-1}/*strA-strB/dfrA-1/tetA/sul-1/sul-2* or *strA-strB/tetA/sul-2* (Figure 6.4). It was noted that both *dfrA-1* and *sul-1* appear to be co-acquired and in rarer cases co-acquired with *bla*_{TEM-1} (Figure 6.3 and 6.4). Within this lineage the AMR determinants previously mentioned appear to be co-located on a large (39.4 kbp) fragment inserted directly 5' the integrase of the LEE pathogenicity island (Figure 6.3).

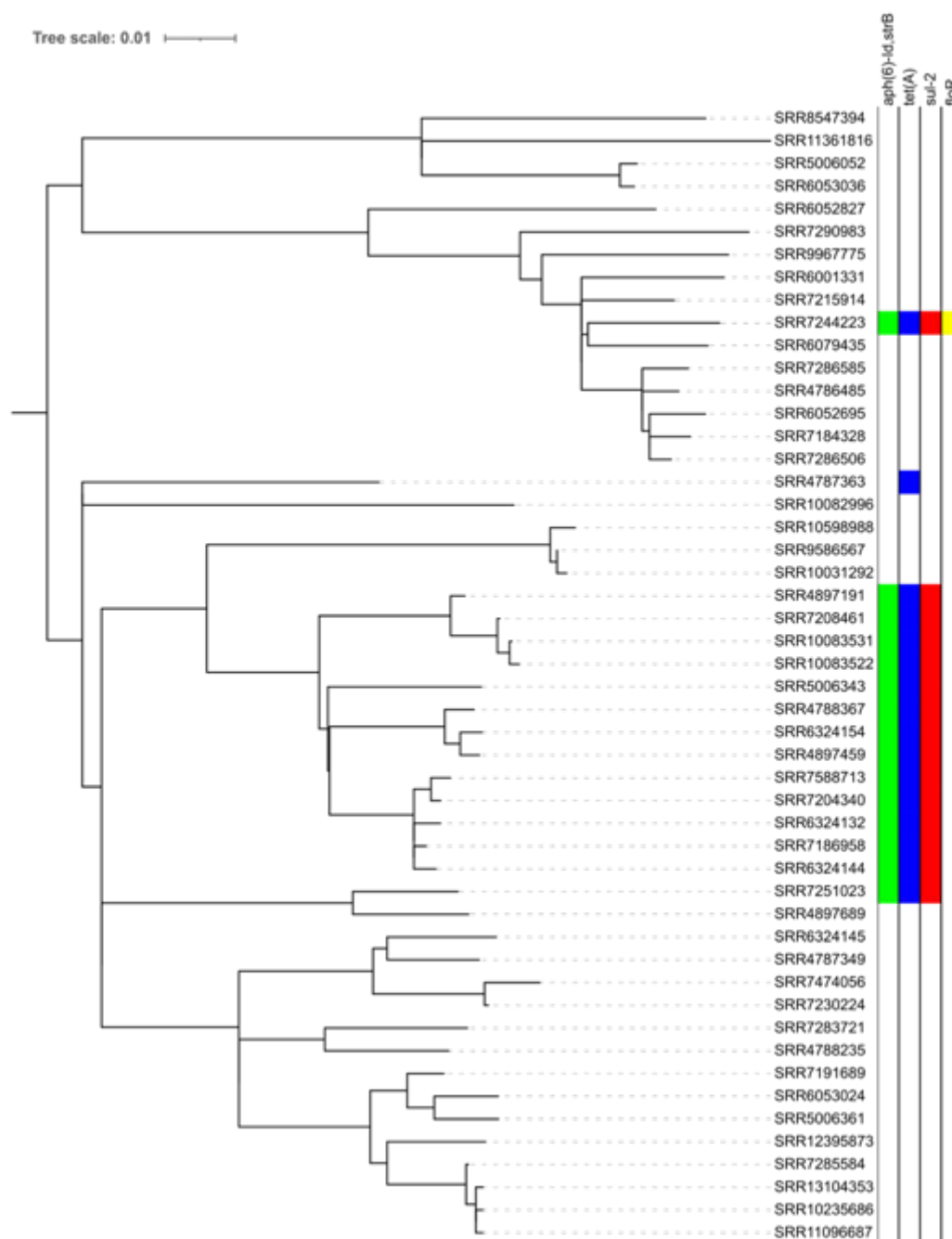


Figure 6.1 – STEC O157:H7 lineage I/II maximum likelihood phylogenetic tree showing the distribution of AMR determinants including *tetA*, *sul2*, *floR* and *aph(6)-Id, strB*.

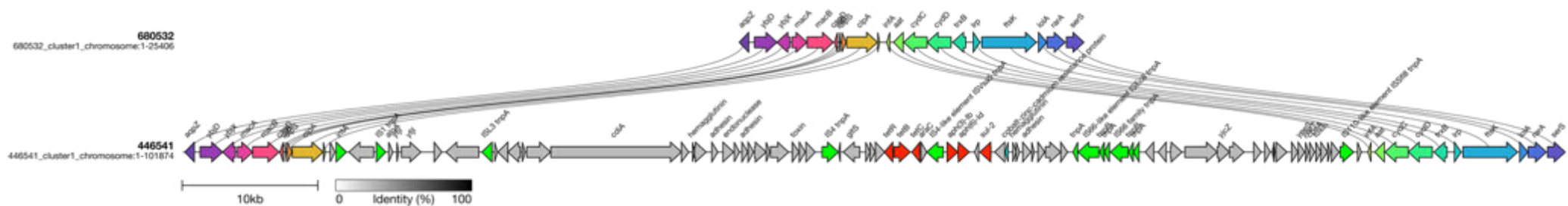


Figure 6.2 – Pairwise clinker alignment of the sub-lineage I/II sample 446541 in which a 9kbp Tn10 intergron was located inside a prophage (located at *clpA*) on the chromosome (at positions 2,080,998 - 2,089,926) containing the AMR determinants *strA*, *strB*, *tetA* and *sul2* (AMR genes shown in red). Black links denote genes of $\geq 90\%$ sequence similarity.

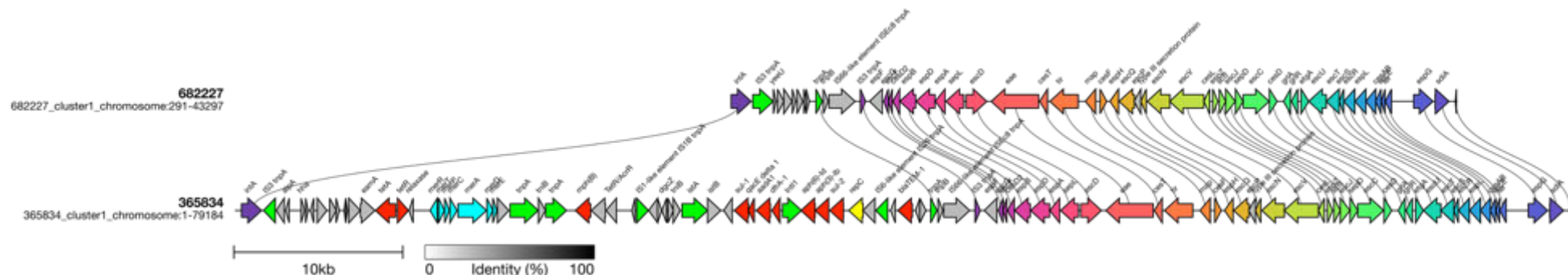


Figure 6.3 – Pairwise clinker alignment of sub-lineage Ib showing that all the AMR determinants (*strA-strB/tetA/sul-2* or *bla_{TEM-1}/strA-strB/dfrA-1/tetA/sul-1/sul-2*) were co-located on the chromosome upon a 39,466bp fragment of DNA inserted within the start of Locus of Enterocyte Effacement (AMR genes shown in red). Black links denote genes of $\geq 90\%$ sequence similarity.

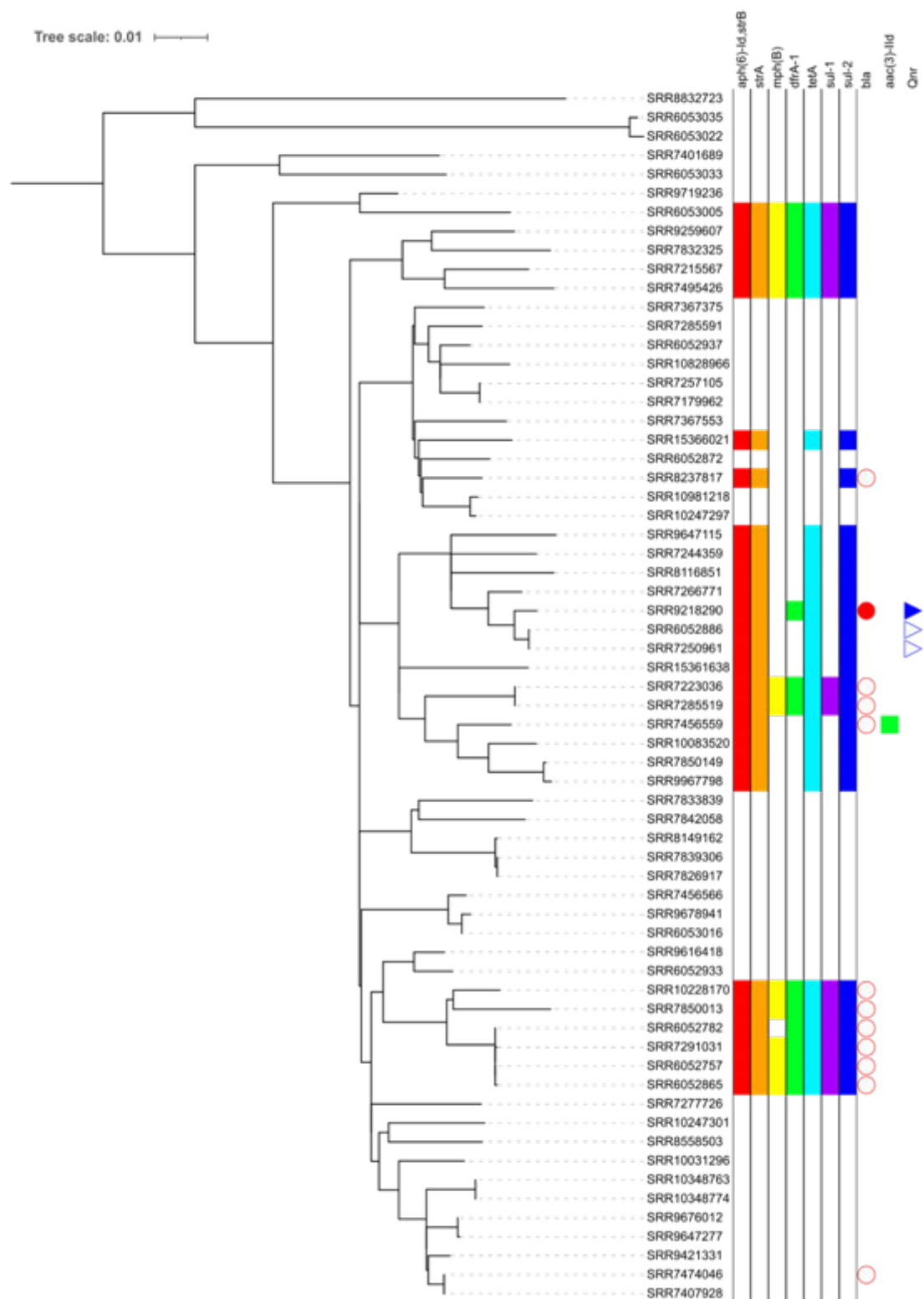


Figure 6.4 – STEC O157:H7 sub-lineage Ib maximum likelihood phylogenetic tree showing the distribution of AMR determinants. Each column represents presence/absence of AMR determinants. The beta-lactams column shows circles for

presence of an AMR determinant, a white circle represents *bla*_{TEM-1} and a red circle represents *bla*_{CTX-M-15}. The quinolones column shows triangles for presence of an AMR determinant, a white triangle represents *qnrS-1* and a blue triangles represents *qnrS-1* and *qnrB-19*.

6.3.5 Phylogenetic analysis of AMR determinants and analysis of Nanopore-based data in sub-lineage IIc.

Sub-lineage IIc showed the most variation of AMR determinants and this was observed throughout the phylogeny of sub-lineage IIc (Figure 6.5). However, there were four different 50-SNP single linkage clusters which contained most of the MDR isolates (Figure 6.5). From here onwards, each of the different 50 SNP single linkage will be known as clusters 1 to 4.

Cluster 1 MDR isolates all shared the same AMR profile of *catA-1*, *floR*, *sul-2*, *dfrA-12*, *aac(3)-IId*, *aadA-2*, *aadA-8b*, *strA/strB* and *bla*_{TEM-1} (Figure 6.5). All the AMR determinants listed appeared to be co-located on a 30,071bp integrative element on the chromosome. This element also included a mercury resistance cassette (Figure 6.6).

Cluster 2 MDR isolates showed a similar resistance profile to cluster 1 isolates however, in this case they were primarily located on an IncHII plasmid with some isolates, located on the chromosome. The observed AMR determinants in this cluster were *floR*, *tetA*, *sul-1*, *sul-2*, *dfrA-1*, *aadD*, *aac(3)-Iva*, *aadA-1b*, *strA/strB* and *bla*_{TEM-1} (Figures 6.5 and 6.6).

Cluster 3 MDR isolates showed no AMR determinants on the chromosome, but genes harboured by IncHII and IncFII plasmids with each plasmid type showing two

distinct profiles. The first profile was *bla*_{TEM-1}, *aadA-1*, *strA/strB*, *sul-1*, *dfrA-1*, *sul-2*, *tetA* and the second profile was *bla*_{CTX-M-15}, *floR*, *sul-2*, *qnrS-1* and *strA/strB* (Figures 6.5 and 6.6).

Finally, Cluster 4 showed the most diversity some isolates harbouring an integrative element (26.8kbp) again integrated at the LEE containing the following AMR determinants *bla*_{TEM-1}, *strA/strB*, *dfrA-8*, *sul-2* and *tetA*. A small subset of isolates also had an IncHI2 plasmid encoding AMR determinates within three different regions of the plasmid. Region1 contained *bla*_{TEM-1}, *aph(3)-Ia*, *aph(6)-Id*, *strA* and *sul-2*. Region2 contained *sul-1*, *aadA-8b*, *dfrA-1* and *tetA*. Region 3 contains *dfrA-12*, *aadA-2*, *aadA-1*. AMR determinants *ermB* and *mph-B* were located on the plasmid but outside the three previously mentioned regions. (Figures 6.5 and 6.6).

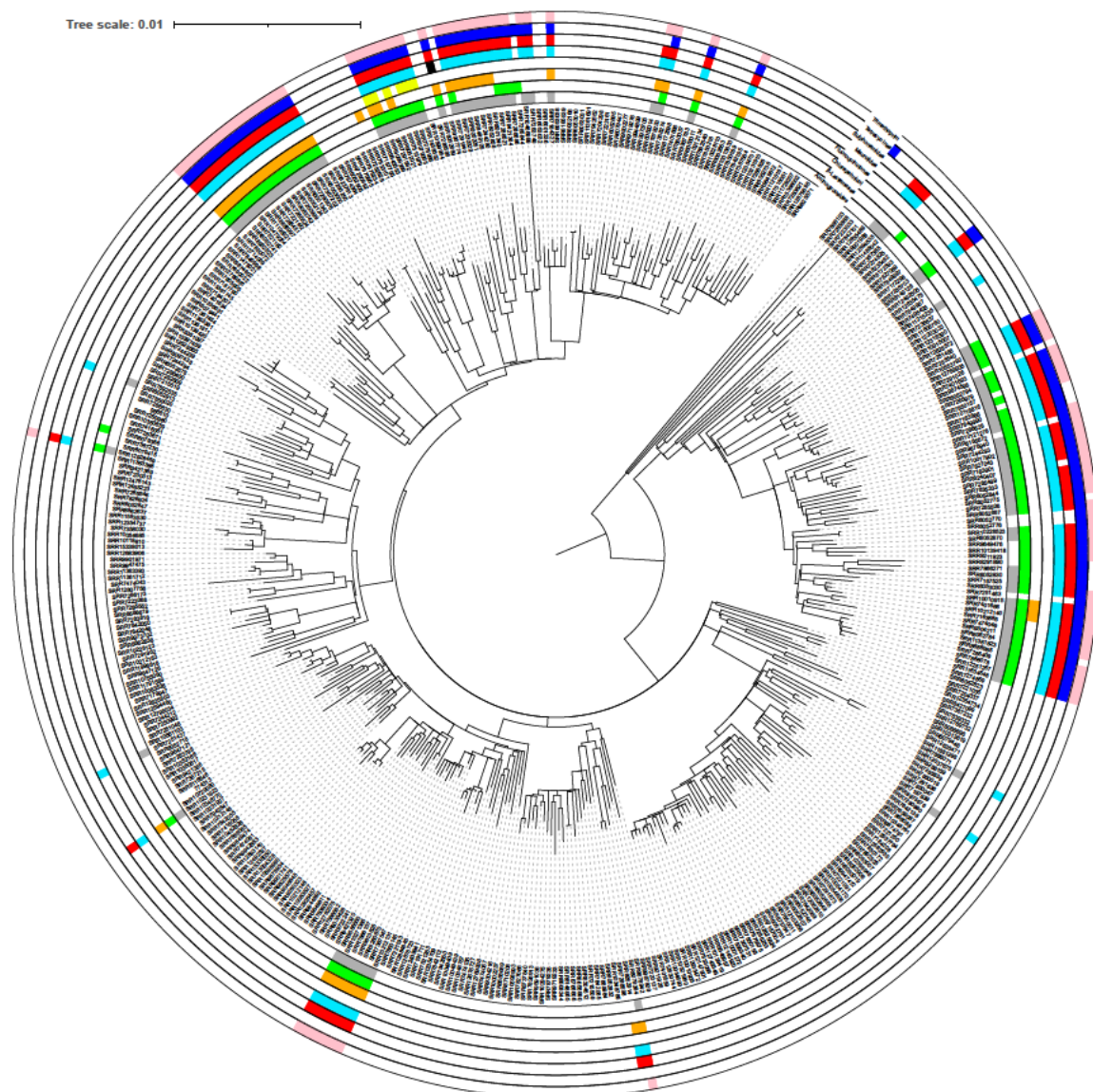


Figure 6.5 – STEC O157:H7 sub-lineage IIc maximum likelihood phylogenetic tree showing the distribution of AMR determinants. Each ring denotes presence/absence of at least one AMR determinants that encode resistance to trimethoprim, tetracycline, sulphonamides, macrolides, fluroquinolones, chloramphenicol, beta-lactams and aminoglycosides (moving in order from outer rings to inner rings).

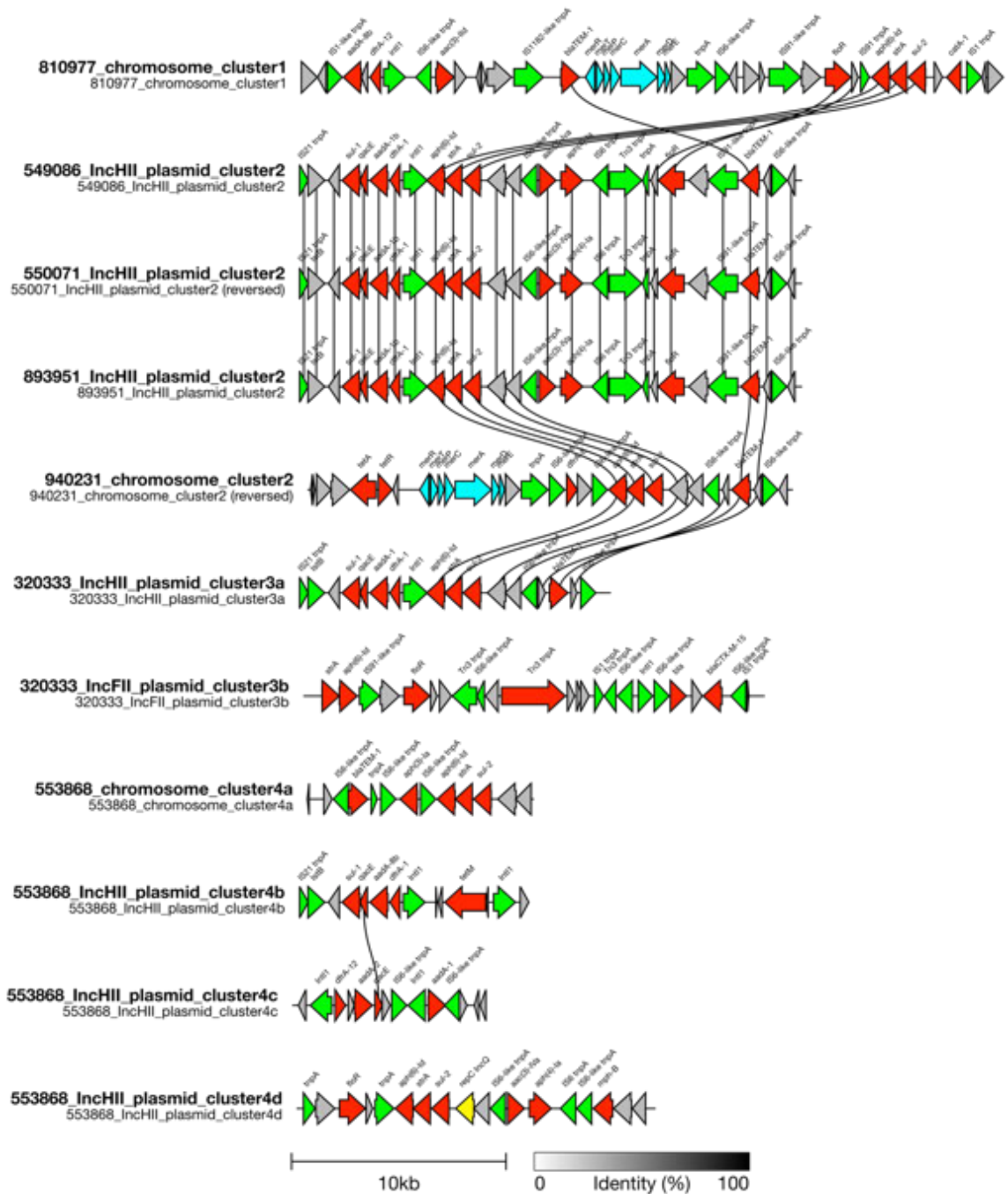


Figure 6.6 – Pairwise alignment of sub-lineage IIc co-located AMR regions. AMR genes shown in red, mobilisation genes shown in green, other resistance positions such as heavy metal resistance shown in blue and plasmid replicons shown in yellow. Black links denote genes of $\geq 90\%$ sequence similarity.

6.4. Chapter 6 discussion.

National surveillance of STEC O157:H7 in England has been in progress since the 1990's. An important aspect of this surveillance is detection and monitoring of AMR profiles in pathogens to understand the transmission risks from the food chain to human infection [228,231,232]. This chapter characterised the AMR gene profiles per sub-lineage of STEC O157:H7. The main observation is that the AMR determinants are unevenly distributed across the STEC O157:H7 population and within sub-lineages. This chapter has demonstrated that sub-lineages Ib, I/II and IIc contained a higher proportion of isolates predicted to be resistant to at least one class of antibiotic, unlike sub-lineages Ia, Ic, IIa and IIb. Previous studies have shown that the dominant strain of STEC O157:H7 in the UK belonged to sub-lineage I/II during the 1980s and 1990s [80,170]. It was noted that sub-lineage I/II contains a higher proportion of isolates showing genotypic resistance compared to sub-lineages Ic or IIb. As a strain replacement event occurred with the dominant strain in the late 1990's belonging to sub-lineage Ic replacing that from sub-lineage I/II could explain the observed decrease in the total number of strains harbouring genotypic resistance [228,231]. Tetracycline, streptomycin and sulphonamide resistance were commonly observed in sub-lineage I/II and is consistent with veterinary therapeutic practices at the time when this sub-lineage was dominant in UK cattle [228]. With increasing controls being placed on the agricultural industry in the late 1990s has likely resulted in the susceptibility observed more recent emerging sub-lineages Ic and IIb. Another possible explanation is that isolates from sub-lineages Ic and IIb could be more susceptible to acquiring foreign DNA harbouring AMR determinants and maintaining these elements.

Sub-lineage Ia appears to harbour a higher proportion of isolates showing genotypic resistance when compared to sub-lineage Ib. It remains unclear at the current time why this is the case despite both sub-lineages being widely distributed across the world.

It has previously been shown that within sub-lineages IIc and IIa, there are clades associated with being endemic in UK cattle and those that are associated with import into the UK via food items and travellers [98,99]. Again, despite sub-lineage IIc harbouring more isolates encoding AMR determinants compared to sub-lineage IIa, the reasons for this are still unclear. There could be a biological reason for this such as certain sub-lineages being able to acquire mobile genetic elements harbouring AMR genes or perhaps an epidemiological reason whereby certain strains are acquired initially in countries where there is less regulated antibiotic use in animal or human treatment.

It is important to note that between sub-lineages of STEC O157:H7 AMR genes were unequally distributed. This also applied within the sub-lineages adding to previous observations [99]. Dallman *et al* 2022 [99], with the conclusion that sub-lineage IIc contains domestic clades which are predominantly susceptible contrary to non-domestic clades which support a high proportion of resistant isolates and perhaps once stabilised these clades may persist in certain environments where antibiotic use is high.

In this chapter, it was observed that genes conferring resistance to the third-generation cephalosporins and azithromycin were plasmid-encoded and therefore had the potential to be transferred to other bacteria in the gut [233,234,235]. These mobile genetic elements hold the potential to transfer from pathogen to commensal

bacteria in the human gut. This ultimately increases the risk for causing life-threatening infections.

Throughout the sub-lineages where Nanopore sequencing took place (I/II, Ib and IIc), it was observed in all sub-lineages integrative elements harbouring AMR genes integrating into the chromosome at prophage or prophage-like regions. This raises a question as to if prophage and prophage-like regions provide hotspots for other non-bacteriophage integrative elements.

Future developments of long-read sequencing technologies combined with real-time WGS-based surveillance will increase the ability to monitor the emergence of MDR strains of STEC O157:H7 and their transmission from the food chain to human infections. The application of long-read sequencing technologies will also allow for the surveillance of mobile genetic elements in their entirety and not simply the presence/absence of AMR resistance determinants.

6.5. Chapter 6 conclusions.

To conclude, UK domestic STEC O157:H7 appear to be susceptible to most classes of antibiotics. This contrasts with travel associated strains of STEC O157:H7 which are more likely to be multi-drug resistant which can exhibit resistance to chloramphenicol, ampicillin and trimethoprim.

With increasing accuracy development with Nanopore sequencing aimed for use of routine surveillance will enable us to monitor the emergence and spread of not just AMR genes within the STEC O157:H7 population but also the mobile genetic elements in which these acquired resistance genes reside. This will enable us to track the transmission of AMR at the pathogen and mobile genetic element level.

Chapter 7:

Discussion/Conclusions

7.1. Final discussion.

Undertaking this PhD project on a part-time basis while working at UKHSA has provided some useful advantages, the biggest being access to a large set of short-read sequencing data which is produced routinely by the gastrointestinal bacteria reference unit (GBRU). However, this thesis aimed to describe both the advantages and challenges of using Nanopore sequencing to investigate outbreaks of STEC O157:H7 especially within outbreaks.

This project was to build upon previous works completed by Shaaban *et al* 2016 ^[81] and Cowley *et al* 2016 ^[139]. The first used a different method of long-read sequencing (Pacific Biosciences) to demonstrate that *stx2a*-encoding prophages were more diverse than *stx2c*-encoding prophages and supported the hypothesis that the *stx2c*-encoding prophage was ancestrally acquired and maintained when compared to other *stx*-encoding prophages. The second was the first to use Nanopore sequencing to compare two STEC O157:H7 outbreaks and demonstrated that the two outbreaks strains had undergone accessory genome changes including plasmid and prophage content relative to each other.

The primary aim of this thesis was to better understand the prophage content that existed when highly pathogenic sub-lineages emerged in the UK and its variation at the outbreak level. This is to highlight markers and other aspects of the STEC O157:H7 accessory genome (such as prophage content) to act as an early warning system to detect emerging highly pathogenic threats in the future. To achieve this, it was crucial to integrate Nanopore sequencing to supplement sequencing methodologies currently in use to be able to detect prophage regions accurately and

develop standardised methods to characterise them. A complete set of results are detailed in previous chapters and the main findings are discussed below.

Due to difficulty resolving the accessory genome through short-read sequencing, the accessory genome of STEC O157:H7 is largely unexplored. The first chapter of this thesis was to design a standardised bioinformatics methodology to process, analyse and characterise long-read sequencing data in order to generate complete genomes of STEC O157:H7 which facilitates the comparison of accessory genome content, including prophage content. This involved two main components, firstly optimising and validating the most accurate methods to generate complete genomes of STEC O157:H7 via *de novo* assembly using Nanopore-based whole genome sequencing data. Secondly, to optimise the alignment-based methods for variant calling and SNP typing using Nanopore sequencing data in order to compare STEC O157:H7 strain relatedness.

There have been many benchmarking studies employed across different software components and versions of tools for processing Nanopore-based sequencing data, including changes in pore type [236], basecalling methods [200], *de novo* assembly [237] and methylation detection [238]. However, there is no standardised validation of the use of Nanopore reference-based variant calling. Reference-based variant calling is the method employed to perform strain relatedness of gastrointestinal pathogens, including STEC O157:H7, in the UK [94,98]. In order to compare and ultimately integrate short and long-read sequencing data, it is essential to be able to determine strain relatedness with Nanopore sequencing using an equivalent approach to the

method used with Illumina-based sequencing at the national level for public health surveillance.

When this PhD project started, the per-read error rate of Nanopore sequencing was around 10% to 15% which was a major limitation of the technology ^[172,173,174,175]. If variant calling and clustering using Nanopore reads was to be performed in conjunction with Illumina reads, accounting for Nanopore per-read accuracy was a critical step.

Two STEC O157:H7 isolates were sequenced using both Nanopore and Illumina platforms and it was possible to compare them to a large database of 4,475 STEC O157:H7 Illumina derived genome sequences. When comparing the Nanopore and Illumina SNPs the majority of discrepant SNPs (those that call a variant or SNP in the alignment of one read set but not the other) were identified in the Nanopore data and not the Illumina data. However, most of these discrepant SNPs were noted to be located in 5-methylcytosine motif sequences, CC(A/T)GG which were likely produced during the basecalling process. Upon accounting for these positions, placement of the Nanopore and Illumina equivalent of each sample upon the same branch, with limited discrepant SNPs (≤ 7 SNPs) was possible.

To generate the most accurate *de novo* assemblies possible a comparison of several publicly available bioinformatics tools designed for assembly correction were compared on a re-sequenced STEC O157:H7 reference genome. Typically, it was noted that short-read polishing post long-read polishing produced a more accurate assembly. When comparing the long-read correction tools, if a trained model can be applied to Medaka it outperforms Nanopolish, however, Nanopolish would be the better method if no such Medaka model existed. The most accurate method post-

optimisation led to a final accuracy of 99.9935% (0.0064% error) which equates to approximately 200 bases difference when compared to STEC O157:H7 reference genome. This level of nucleotide accuracy was newly achieved at the time as a result of the optimisation of the assembly correction process. The remaining discrepancies appeared in paralogous sequences and prophage regions of the STEC genome. Using short-read correction at the end of the workflow led to more errors being corrected into the assembly due to ambiguity when aligning short reads to the prophage regions. Of the 200 remaining differences, 95.5% of these were determined to be within prophage regions, highlighting the need for specific corrections methods particularly in homologous regions of the STEC O157:H7 genome. This demonstrates the need to consider the homologous regions of the STEC O157:H7 genome or assembly correction using short-read data could be detrimental to prophage content by correcting in erroneous sequences due to ambiguous alignment.

Currently, both short- and long-read technologies are used for public health surveillance, and there is a need to integrate the outputs so that all the data can be analysed in the same way. Continuous updates to the library preparation chemistry and basecalling software tools will improve robustness of variant calling and assembly correction thus enabling Nanopore to compete with short-read platforms in terms of per-read quality, and ultimately enabling the two technologies to be used interchangeably in clinical and public health settings. Completing this optimisation process provided a robust foundation, facilitating the characterisation of STEC O157:H7 genomes in a standardised way for the remainder of the PhD project.

With a standardised method for the generation of complete STEC O157:H7 genomes and for determined reference-based strain relatedness, the next step was to compare *stx*-encoding prophages from the initial emergence of high-pathogenic lineages of STEC O157:H7 in the UK. This is important as it provided an initial understanding of the prophage content of highly pathogenic STEC genomes.

Understanding how highly pathogenic lineages emerge allows us to apply methods to search for early warning markers during surveillance, which highlights emerging threats to public health.

Three different *stx1a*-encoding prophages were observed within two different sub-lineages (Ia and IIc) of the strains of STEC O157:H7 characterised. The observed *stx1a*-encoding prophages were structurally similar despite differences in geography and time of the isolated hosts. This corroborates the findings previously described that *stx1a*-encoding prophages were likely acquired ancestrally and vertically inherited throughout the population [80,81].

Similarly, *stx2c*-encoding bacteriophages characterised showed little variation and like *stx1a*-encoding prophages and supports the hypothesis that it was acquired ancestrally and maintained by vertical inheritance [80,81]. This is particularly highlighted by the evidence that *stx2c*-encoding prophages from hosts isolated in 1983 and 2016 shared a high sequence similarity.

Prophages encoding the *stx2a* subtype showed a much higher level of genetic diversity, compared to the level of diversity between *stx2c*- and *stx1a*-encoding prophages. Levels of diversity were high even between *stx2a* encoding prophages from strains that were temporally and/or geographically related. During the 1980s, in the UK, STEC O157:H7 HUS-associated strains were typically part of sub-lineages

Ia and I/IIa (primarily PT4 and PT1) [18,35]. By the end of the 1980s, notifications of cases infected with strains associated with these sub-lineages declined and were replaced by cases infected with an emerging PT49 I/IIb strain. This analysis has demonstrated the *stx2a*-encoding prophage in the emergent PT49 I/IIb was structurally very different from *stx2a*-encoding prophages from PT4, PT1 sub-lineages Ia and I/IIa indicating the acquisition of the *stx2a*-encoding prophage into the PT49 I/IIb from a previously unknown reservoir. In the 1990s, a sub-lineage Ic strain circulating through UK and Irish cattle populations acquired a *stx2a*-encoding prophage (SBI is *argW*) changing the phage type from PT32 to PT21/28 and marked an increase in pathogenicity [80,81]. This strain replaced the previous sub-lineage Ia, I/IIa and I/IIb population as the most common STEC O157:H7 detected in clinical cases in the UK. More recently, the sub-lineage Ic PT21/28s infections in humans slowly declined with an increase in an emerging IIb PT8 strain, once again, linked to acquisition of a *stx2a*-encoding prophage [100]. The *stx2a*-encoding prophage in sub-lineage IIb is extremely similar to the *stx2c*-encoding prophage structure and the *stx2a*-encoding prophage was located at *sbcB*, again a site usually associated with *stx2c*-encoding prophages. This most likely to have occurred through horizontal exchange of the *stx2a* gene module into an existing *stx2c*-encoding prophage, driven by homologous recombination between very similar prophage sequences, in this case the prophage structural genes in the *stx2c*- and *stx2a*-encoding prophages. The analyses in this thesis also provided evidence that a dynamic environmental reservoir of *stx2a*-encoding prophages exists and most likely circulates under the radar of public health surveillance systems. The ability of *stx*-encoding prophages protentional to integrate into domestic host of STEC O157:H7, and other STEC

serotypes, and drive a more severe infection poses a significant threat to public health. Going forward it will be important to monitor STEC host strains and their *stx*-encoding prophage complement, as this can provide an early warning to emergence of high-pathogenic sub-lineages.

With a reference set of *stx*-encoding prophages in UK high-pathogenic sub-lineages characterised, the next phase of this project was to explore whether there was any previously hidden genome variation of STEC O157:H7 at the intra-outbreak strain level. To achieve this, four historical outbreaks of STEC O157:H7 associated with different vehicles and transmission routes were sequenced retrospectively using Nanopore platforms. One outbreak was associated with handling contaminated raw dog food, another was caused by environmental contamination at a mud-based obstacle course, the third was linked to consumption of raw drinking milk and the fourth was associated with direct animal contact at a petting farm.

When comparing strains using SNP-typing regardless of the outbreak analysed, there were systemic differences between the Illumina-based results and Nanopore-based results. Both technologies produced errors, homopolymer-based basecalling errors from the Nanopore data ^[182,201] and ambiguous alignment to homologous (in particular prophage regions) and paralogous regions of reference genome for the Illumina data ^[202] led to discrepant variant calls in each case. However, despite these errors, both datasets confirmed the close genetic relationship for all samples in each of the outbreaks.

In each outbreak, the prophage content was compared to a well described reference genome ^[75] which belonged to sub-lineage Ia. Each of the four outbreaks belonged

to either sub-lineage Ic or IIb, so are genetically distinct from the reference genome. The outbreaks all occurred after 2010 so are also temporally distinct and geographically distinct from reference strain which was originally isolated in Japan in 1996. Despite this the prophage commonality shared between the strain Sakai and the outbreak strains indicates stability of some (approximately 40%) the non-*stx*-encoding prophage content over time and space. When comparing prophage content between host strains that are closer genetically for example, from the same sub-lineage, there was an increase in similarity of prophage content. As a benchmark for prophage content within STEC O157:H7 across the population does not yet exist, is difficult to elucidate if these changes in content have any impact on the host's ability to survive or pathogenicity.

Comparing *stx*-encoding prophages specifically there was much more diversity observed and in the case of the outbreak associated with raw drinking milk consumption, there were two copies of a *stx2a*-encoding prophage, in the genomes of the outbreak strain, the second located at an integration site (*sbcB*) associated with and structurally similar to *stx2c*-encoding prophages. Similarly, the outbreak associated with environmental contamination also had a single *stx2a*-encoding prophage but located at *sbcB* and structurally like *stx2c*-encoding prophages. Both the effect of the more pathogenic *stx2a* gene in a *stx2c*-encoding prophage on expression of the toxin and the effect of harbouring two or more of the same Shiga toxin subtype are yet to be determined.

Large chromosomal rearrangements (LCR) were found in three of the outbreaks. The first associated with raw drinking milk had an inversion with a second nested inversion in a single sample that was isolated from the cattle in which the raw milk

was produced. Secondly, the temporally distant samples in the environmental exposure outbreak had a large inversion relative to the other samples. Thirdly, the outbreak associated with direct animal contact showed much more LCR than the other outbreaks, with multiple inversions observed and in one case an inverted translocation affecting the *stx*-encoding prophage.

The effect of these LCRs on the pathogenicity or the ability of the pathogen to adapt in the reservoir or infectious states have not been determined however, it is likely driving the genetic diversification of prophages as genetic material from each prophage at the end of the inversions is exchanged. My work from this chapter highlights a higher level of genome variation in samples within an outbreak that has previously been observed. In each of the outbreaks there was genome variation in prophage and plasmid content. Similarly, LCRs observed within outbreaks are also a novel finding.

With the development of long-read sequencing technology, we can now detect and describe both large- and small-scale structural variation and explore the effect on phenotype. The ability to characterise the accessory genome in this format is the first step to understanding the significance of these micro-evolutionary events and their impact on the evolutionary history, virulence, and potentially the likely source and transmission of this zoonotic, foodborne pathogen.

The previous chapter highlighted that large chromosomal rearrangements (LCRs) occur within outbreaks of STEC O157:H7. Focusing on these events specifically, lambdoid prophages of the STEC O157:H7 chromosome were determined to be the boundaries in all occurrences observed. Searching for homologous sequences

≥2.5kbp throughout the chromosome of STEC O157:H7 highlighted many of these regions of with majority of them existing within lambdoid prophages. Lambdoid prophages share the same structural genes (head, tail fibres, sheath etc) [111,120,217]. All observed LCRs occurred with breakpoints within 1Mbp either side of the terminus of replication of the chromosome meaning the homologous regions are spatially closer together than other regions of the chromosome. This combined with previous observations [207] corroborates the likely mechanism being homologous recombination due to double-stranded breaks during the SOS response which is responsible for both Shiga-toxin expression from *stx*-encoding prophages and triggers recombinational-based DNA repair [120]. LCRs are also likely to drive genetic diversification of prophages with modules being exchanged at the breakpoints of the LCR. The effects of the LCR on the prophages ability to excise and function remains undetermined.

The existence of isoforms of the same chromosome within a single culture of STEC O157:H7 was a novel finding. Where a single long-read dataset contained reads in support of different isoforms of the same chromosome typically where an LCR has occurred. The existence of both isoforms in a single sample read sets suggests that within a single pure culture, cells contain both chromosomal isoforms. This could have implications for bioinformatics processed such as *de novo* assembly.

Generating complete assemblies of STEC O157:H7 using long-read sequencing also provides an opportunity to characterise the mobile genetic elements associated with antimicrobial resistance genes. Combining this with national surveillance whole-

genome sequencing data allowed for a description of AMR in the contemporary STEC O157:H7 UK population.

Compared to other gastrointestinal pathogens, such as *Shigella* species and the typhoidal salmonellae, the frequency of AMR genes within the population is less frequent in STEC O157:H7, perhaps because AMR therapy for STEC infection is contraindicated. There was an unequal distribution of AMR carriage across the sub-lineages of STEC O157:H7 with sub-lineages Ib, I/II and IIc containing a higher proportion of isolates showing genotypic resistance to at least one class of antibiotic, unlike sub-lineages Ia, Ic, IIa and IIb. Evidence of increasing controls in the agricultural industry during the 1990s was observed as streptomycin, tetracycline and sulphonamide resistance (consistent with veterinary therapeutic practices) was typically found in sub-lineage I/II which was the dominant UK strain in the 1980s [228]. The susceptibility observed in more recent emerging sub-lineages Ic and IIb is most likely due to these imposed controls. Sub-lineages IIc and IIa also showed a higher rate of AMR genes detected as these sub-lineages are both associated with travel outside the UK and food imports [98,99].

A novel finding of this chapter was the observed integration of mobile genetic elements within prophage and prophage-like regions of the STEC O157:H7 chromosome. A wider long-read sequencing study including a larger sample number can highlight the propensity for this observed phenomenon.

Combining the rapid development of long-read sequencing technologies with real-time WGS-based surveillance will increase the ability to monitor the emergence of MDR strains of STEC O157:H7 and their transmission from food chain to human infections. The ability to monitor mobile genetic elements in their entirety and not just

the presence or absence of AMR genes is required for a better understanding of the spread of AMR within the STEC O157:H7 population. These mobile genetic elements have the capacity to transfer from pathogen to commensal bacteria in the human gut, thus facilitating the spread of AMR and the risk of life-threatening infections.

7.2. Limitations of analyses.

At the time this study was carried out, long read sequencing was low throughput and relatively expensive compared to short read Illumina sequencing, and thus the number of samples and the number of outbreaks analysed were small. Specifically, the outbreaks fell within two sub-lineages, Ic and IIb, meaning the full diversity of the STEC O157:H7 population is not captured. It is likely that the intra-outbreak accessory genome variation observed also occurs in other sub-lineages of STEC O157:H7, but this needs to be confirmed.

Similarly, this means only the prophage diversity associated with those sub-lineages Ic and IIb was observed however, there is likely to be more prophage diversity to be uncovered within the untested sub-lineages. Despite this, sub-lineages where high-pathogenic strains emerged were specifically targeted as they pose the greatest public health threat

Each set of outbreak samples underwent Nanopore-based whole genome sequencing at different times. As a result, some of the basecalling software produced changes in per-read quality, although the bioinformatics pipeline was designed to account for this variability as much as possible. Nanopore-based chemistries (library preparation), sequencing (type of pores used on the flow cells)

and software (basecalling) have improved dramatically during the timeframe to complete this work. As all Nanopore sequencing were performed in the early phase of these works the improvements were not utilised however, given the nature of Nanopore sequencing where the raw signal is stored digitally, the option to utilise the newer versions of basecalling software on the raw signal for re-analysis is achievable.

7.3. Proposed future works.

There are several aspects of the work produced that can be continued to either generate new context, reduce limitations and investigating underlying biological mechanisms behind events that have been observed.

Firstly, a wider analysis of the population of STEC O157:H7 should be considered. Representatives from other sub-lineages such as Ia, Ib, I/IIa, IIa and IIc should be selected, this would allow for a reference set of prophages to be characterised from the UK domestic population of STEC O157:H7. This reference set could allow for phylogeographic analyses to be developed for determination if the host strain is of domestic or non-domestic origin. Secondly this reference set will allow for a more complete analysis of prophage content by allowing for characterisation and comparison of prophage content on a per-sub-lineage basis.

The results produced in this thesis is primary bioinformatics based as a result it is lacking *in vitro* and *in vivo* confirmation. This work would benefit a collaboration with a laboratory to address some of the questions raised. An example of this would be to further investigate the underlying biological mechanism responsible for LCR formation. This could be achieved by several methods, firstly using optical mapping

approaches to capture all LCRs within a culture of STEC O157:H7. As optical mapping has a higher resolution to detect low level variation it can highlight LCRs that exist but are missed during long-read sequencing assembly methods. Another would be to inhibit *RecBCD* functionality within STEC O157:H7 cells in order to understand if double strand breaks that lead to LCR formation are due to the host DNA repair mechanisms or the prophage DNA repair mechanisms. This combined with targeted long-read sequencing could also detect where double strand breaks occur and if this is randomised across the chromosome or if breaks occur in particular regions of the STEC O157:H7 chromosome.

Another proposed piece of work is confirming if *in vitro* culture of STEC O157:H7 induces LCR formation. Long-read sequencing of a repeated set of strains under various storage conditions (time, temperature, storage media) will highlight any storage conditions that perhaps would artificially induce LCR formation.

Finally, a bioinformatics-based project to containerise and package the bioinformatics workflows for the analysis of long-read sequencing data specific to STEC genome characterisation. The stand-alone software should be optimised and validated and can be of use in public health microbiological analyses.

7.4. Final conclusions.

This thesis has demonstrated that the genome of STEC O157:H7 is more dynamic than previously understood with variation occurring in prophage content, plasmid content and large chromosomal rearrangements occurring all within the outbreak level. Long-read Nanopore sequencing has been highlighted as a useful technology to detect and characterise accessory genome variation of STEC O157:H7.

Supplementing high-throughput sequencing for public health surveillance with long-read sequencing can enable us to monitor the emergence of highly pathogenic sub-lineages within the food chain and provide accurate risk assessments during outbreak investigations of this zoonotic foodborne pathogen.

Reference List/Bibliography:

1. Shulman ST, Friedmann HC, Sims RH. Theodor Escherich. The First Pediatric Infectious Diseases Physician? *Clin Infect Dis*. 2007;45(8):1025–9. doi: 10.1086/521946. PMID 17879920.
2. Leimbach A, Hacker J, Dobrindt U. *E. coli* as an All-Rounder: The Thin Line Between Commensalism and Pathogenicity. *Curr Top Microbiol Immunol*. 2013;358:3-32. doi: 10.1007/82_2012_303. PMID 23340801.
3. Kaper JB, Nataro JP, Mobley HLT. Pathogenic *Escherichia coli*. *Nat Rev Microbiol*. 2004;2(2):123-40. doi: 10.1038/nrmicro818. PMID 15040260.
4. Nataro JP, Kaper JB. Diarrheagenic *Escherichia coli*. *Clin Microbiol Rev*. 1998;11(1):142–201. doi: 10.1128/CMR.11.1.142. PMID 9457432.
5. Czeczulin JR, Balepur S, Hicks S, Phillips A, Hall R, Kothary MH, Navarro-Garcia F, Nataro JP. Aggregative adherence fimbria II, a second fimbrial antigen mediating aggregative adherence in enteroaggregative *Escherichia coli*. *Infect Immun*. 1997;65(10):4135–45. doi: 10.1128/iai.65.10.4135-4145.1997. PMID 9317019.
6. Nataro JP, Yikang D, Yingkang D, Walker K. *AggR*, a transcriptional activator of aggregative adherence fimbria I expression in enteroaggregative *Escherichia coli*. *J Bacteriol*. 1994;176(15):4691–9. doi: 10.1128/jb.176.15.4691-4699.1994. PMID 7913930.
7. Spangler BD. Structure and function of cholera toxin and the related *Escherichia coli* heat-labile enterotoxin. *Microbiol Rev*. 1992;56(4):622–47. doi: 10.1128/mr.56.4.622-647.1992. PMID 1480112.
8. Kunkel SL, Ronertson DC. Purification and chemical characterization of the heat-labile enterotoxin produced by enterotoxigenic *Escherichia coli*. *Infect Immun*. 1979 Aug;25(2):586-96. doi: 10.1128/iai.25.2.586-596.1979. PMID 39893.
9. Pupo GM, Lan R, Reeves PR. Multiple independent origins of *Shigella* clones of *Escherichia coli* and convergent evolution of many of their characteristics. *Proc Natl Acad Sci USA*. 2000;97(19):10567–72. doi: 10.1073/pnas.180094797. PMID 10954745.
10. Sansonetti PJ, Phalipon A, Arondel J, Thirumalai K, Banerjee S, Akira S, Takeda K, Zychinsky A. Caspase-1 activation of IL-1 β and IL-18 are essential for *Shigella flexneri*-induced inflammation. *Immunity*. 2000;12(5):581–90. doi: 10.1016/s1074-7613(00)80209-5. PMID 10843390.
11. Peiffer I, Servin AL, Bernet-Camard MF. Piracy of decay-accelerating factor (CD55) signal transduction by the diffusely adhering strain *Escherichia coli* C1845 promotes cytoskeletal F-actin rearrangements in cultured human intestinal INT407 cells. *Infect Immun*. 1998;66(9):4036–42. doi: 10.1128/IAI.66.9.4036-4042.1998. PMID 9712744.
12. Bernet-Camard MF, Coconnier MH, Hudault S, Servin AL. Pathogenicity of the diffusely adhering strain *Escherichia coli* C1845: F1845 adhesin-decay accelerating factor interaction, brush border microvillus injury, and actin disassembly in cultured human intestinal epithelial cells. *Infect Immun*. 1996;64(6):1918–28. doi: 10.1128/iai.64.6.1918-1928.1996. PMID 8675288.
13. Jerse AE, Yu J, Tall BD, Kaper JB. A genetic locus of enteropathogenic *Escherichia coli* necessary for the production of attaching and effacing lesions on tissue culture cells. *Proc Natl Acad Sci USA*. 1990;87(20):7839–43. doi: 10.1073/pnas.87.20.7839. PMID 2172966.
14. Fakhouri F, Zuber J, Frémeaux-Bacchi V, Loirat C. Haemolytic uraemic syndrome. *Lancet*. 2017;390(10095):681–96. doi: 10.1016/S0140-6736(17)30062-4. PMID 28242109.
15. Iguchi A, Iyoda S, Kikuchi T, Ogura Y, Katsura K, Ohnishi M, Hayashi T, Thomson NR. A complete view of the genetic diversity of the *Escherichia coli* O-antigen biosynthesis gene cluster. *DNA Res*. 2015 Feb;22(1):101-7. doi: 10.1093/dnares/dsu043. PMID 25428893.

16. Riley LW, Remis RS, Helgerson SD, McGee HB, Wells JG, David BR, Hebert RJ, Olcott ES, Johnson LM, Hargrett NT, Blake PA, Cohen ML. Hemorrhagic colitis associated with a rare *Escherichia coli* serotype. *N Engl J Med*. 1983 Mar 24;308(12):681-5. PMID 6338386.
17. DebRoy C, Roberts E, Valadez AM, Dudley EG, Cutter CN. Detection of Shiga toxin-producing *Escherichia coli* O26, O45, O103, O111, O113, O121, O145, and O157 serogroups by multiplex polymerase chain reaction of the *wzx* gene of the O-antigen gene cluster. *Foodborne Pathog Dis*. 2011;8(5):651–2. doi: 10.1089/fpd.2010.0769. PMID 21548768.
18. Scotland SM, Smith HR, Willshaw GA, Rowe B. Vero cytotoxin production in strain of *Escherichia coli* is determined by genes carried on bacteriophage. *Lancet*. 1983;2(8343):216. doi: 10.1016/s0140-6736(83)90192-7. PMID 6135046.
19. Krüger A, Lucchesi PMA. Shiga toxins and *stx* phages: highly diverse entities. *Microbiology*. 2015;161(Pt 3):451–62. doi: 10.1099/mic.0.000003. PMID 25479836.
20. Bai X, Fu S, Zhang J, Fan R, Xu Y, Sun H, He X, Xu J, Xiong Y. Identification and pathogenomic analysis of an *Escherichia coli* strain producing a novel Shiga toxin 2 subtype. *Sci Rep*. 2018 Apr 30;8(1):6756. doi: 10.1038/s41598-018-25233-x. PMID 29712985.
21. Lacher DW, Gangiredla J, Patel I, Elkins CA, Feng PCH. Use of the *Escherichia coli* Identification Microarray for Characterizing the Health Risks of Shiga Toxin-Producing *Escherichia coli* Isolated from Foods. *J Food Prot*. 2016 Oct;79(10):1656-1662. doi: 10.4315/0362-028X.JFP-16-176. PMID 28221838.
22. Stanton, Cebelinski and Wang 01/09/2021 <https://www.ncbi.nlm.nih.gov/nuccore/MZ571121>.
23. Yang X, Bai X, Zhang J, Sun H, Fu S, Fan R, He X, Scheutz F, Matussek A, Xiong Y. *Escherichia coli* strains producing a novel Shiga toxin 2 subtype circulate in China. *Int J Med Microbiol*. 2020 Jan;310(1):151377. doi: 10.1016/j.ijmm.2019.151377. PMID 31757694.
24. EFSA BIOHAZ Panel, Koutsoumanis K, Allende A, Alvarez-Ordóñez A, Bover-Cid S, Chemaly M, Davies R, De Cesare A, Herman L, Hilbert F, Lindqvist R, Nauta M, Peixe L, Ru G, Simmons M, Skandamis P, Suffredini E, Jenkins C, Monteiro Pires S, Morabito S, Niskanen T, Scheutz F, da Silva Felício MT, Messens W and Bolton D, 2020. Scientific Opinion on the pathogenicity assessment of Shiga toxin-producing *Escherichia coli* (STEC) and the public health risk posed by contamination of food with STEC. *EFSA Journal* 2020;18(1):5967. doi: 10.2903/j.efsa.2020.5967.
25. Bai X, Scheutz F, Dahlgren HM, Hedenström I, Jernberg C. Characterization of Clinical *Escherichia coli* Strains Producing a Novel Shiga Toxin 2 Subtype in Sweden and Denmark. *Microorganisms*. 2021 Nov 17;9(11):2374. doi: 10.3390/microorganisms9112374. PMID 34835499.
26. Obrig TG. *Escherichia coli* Shiga Toxin Mechanisms of Action in Renal Disease. *Toxins*. 2010 Dec 2;2(12):2769-2794. doi: 10.3390/toxins2122769. PMID 21297888.
27. Melton-Celsa AR. Shiga Toxin (Stx) Classification, Structure, and Function. *Microbiol Spectr*. 2014;2(4):EHEC-0024-2013. doi: 10.1128/microbiolspec.EHEC-0024-2013. PMID 25530917.
28. Jogannes L, Römer W. Shiga toxins--from cell biology to biomedical applications. *Nat Rev Microbiol*. 2010 Feb;8(2):105-16. doi: 10.1038/nrmicro2279. PMID 20023663.
29. Stevens MP, Frankel GM. The Locus of Enterocyte Effacement and Associated Virulence Factors of Enterohemorrhagic *Escherichia coli*. *Microbiol Spectr*. 2014;2(4):EHEC-0007-2013. doi: 10.1128/microbiolspec.EHEC-0007-2013. PMID 26104209.
30. Kenny B, Finlay BB. Protein secretion by enteropathogenic *Escherichia coli* is essential for transducing signals to epithelial cells. *Proc Natl Acad Sci USA*. 1995 Aug 15;92(17):7991-5. doi: 10.1073/pnas.92.17.7991. PMID 7644526.
31. Kalman D, Weiner OD, Goosney DL, Sedat JW, Finlay BB, Abo A, Bishop JM. Enteropathogenic *E. coli* acts through WASP and Arp2/3 complex to form actin pedestals. *Nat Cell Biol*. 1999 Oct;1(6):389-91. doi: 10.1038/14087. PMID 10559969.

32. Tobe T, Beatson SA, Taniguchi H, Abe H, Bailey CM, Fivian A, Younis R, Matthews S, Marches O, Frankel G, Hayashi T, Pallen MJ. An extensive repertoire of type III secretion effectors in *Escherichia coli* O157 and the role of lambdoid phages in their dissemination. *Proc Natl Acad Sci USA*. 2006;103(40):14941–6. doi: 10.1073/pnas.0604891103. PMID 16990433.
33. Wolfson EB, Elvidge J, Tahoun A, Gilletpie T, Mantell J, McAteer SP, Rossez Y, Paxton E, Lane F, Shaw DJ, Gill AC, Stevens J, Verkade P, Blocker A, Manajan A, Gally DL. The interaction of *Escherichia coli* O157:H7 and *Salmonella* Typhimurium flagella with host cell membranes and cytoskeletal components. *Microbiology*. 2020 Oct;166(10):947-965. doi: 10.1099/mic.0.000959. PMID 32886602.
34. Hagel C, Krasemann S, Löffler J, Püschel K, Magnus T, Glatzel M. Upregulation of Shiga toxin receptor CD77/Gb3 and interleukin-1 β expression in the brain of EHEC patients with hemolytic uremic syndrome and neurologic symptoms. *Brain Pathol*. 2015;25(2):146–56. doi: 10.1111/bpa.12166. PMID 24989888.
35. Taylor CM, White RH, Winterborn MH, Rowe B. Haemolytic-uraemic syndrome: clinical experience of an outbreak in the West Midlands. *Br Med J (Clin Res Ed)*. 1986 Jun 7;292(6534):1513-6. doi: 10.1136/bmj.292.6534.1513. PMID 3087499.
36. Yukioka H, Kurita S. *Escherichia coli* O157 infection disaster in Japan, 1996. *Eur J Emerg Med*. 1997 Sep;4(3):165. doi: 10.1097/00063110-199709000-00009. PMID 9426998.
37. López AG, Zazo JL, Díaz RD, de Guevara C. Incidence of *E. coli* O157:H7 and other enteropathogens in a Spanish hospital. *Eur J Epidemiol*. 2000 Mar;16(3):303-4. doi: 10.1023/a:1007607020841. PMID 10870948.
38. Pennington H. *Escherichia coli* O157. *Lancet*. 2010 Oct 23;376(9750):1428-35. doi: 10.1016/S0140-6736(10)60963-4. PMID 20971366.
39. Tuttle J, Gomez T, Doyle MP, Wells JG, Zhao T, Tauxe RV, Griffin PM. Lessons from a large outbreak of *Escherichia coli* O157:H7 infections: insights into the infectious dose and method of widespread contamination of hamburger patties. *Epidemiol Infect* 1999;122(2):185-92. doi: 10.1017/s0950268898001976. PMID 10355781.
40. Willshaw GA, Thirlwell J, Jones AP, Parry S, Salmon RL, Hickey M. Vero cytotoxin-producing *Escherichia coli* O157 in beefburgers linked to an outbreak of diarrhoea, haemorrhagic colitis and haemolytic uraemic syndrome in Britain. *Lett Appl Microbiol*. 1994;19(5):304-7. doi: 10.1111/j.1472-765x.1994.tb00461.x. PMID 7765442.
41. Naylor SW, Low C, Besser TE, Mahajan A, Gunn GJ, Pearce MC, McKendrick IJ, Smith DGE, Gally DL. Lymphoid follicle-dense mucosa at the terminal rectum is the principal site of colonization of enterohemorrhagic *Escherichia coli* O157:H7 in the bovine host. *Infect Immun*. 2003 Mar;71(3):1505-12. doi: 10.1128/IAI.71.3.1505-1512.2003. PMID 12595469.
42. Ferens WA, Hovde CJ. *Escherichia coli* O157:H7: animal reservoir and sources of human infection. *Foodborne Pathog Dis*. 2011 Apr;8(4):465-87. doi: 10.1089/fpd.2010.0673. PMID 21117940.
43. Chapman PA, Siddons CA, Gerdan Malo AT, Harkin MA. A 1-year study of *Escherichia coli* O157 in cattle, sheep, pigs and poultry. *Epidemiol Infect*. 1997 Oct;119(2):245-50. doi: 10.1017/s0950268897007826. PMID 9363024.
44. Milnes AS, Stewart I, Clifton-Hadley FA, Davies, RH, Newell DG, Sayers AR, Cheasty T, Cassar C, Ridley A, Cook AJC, Evans SJ, Teale CJ, Smith RP, McNally A, Toszeghy M, Futter R, Kay A, Paiba GA. Intestinal carriage of verocytotoxigenic *Escherichia coli* O157, *Salmonella*, thermophilic *Campylobacter* and *Yersinia enterocolitica*, in cattle, sheep and pigs at slaughter in Great Britain during 2003. *Epidemiol Infect*. 2008 Jun;136(6):739-51. doi: 10.1017/S0950268807009223. PMID 17655782.
45. Orden JA, Cortés C, Horcajo P, De la Fuente R, Blanco JE, Mora A, López C, Blanco J, Contreras A, Sánchez A, Corrales JC, Domínguez-Bernal G. A longitudinal study of

- verotoxin-producing *Escherichia coli* in two dairy goat herds. *Vet Microbiol.* 2008 Dec 10;132(3-4):428-34. doi: 10.1016/j.vetmic.2008.05.021. PMID 18603384.
46. Hancock DD, Besser TE, Rice DH, Ebel ED, Herriott DE, Carpenter LV. Multiple sources of *Escherichia coli* O157 in feedlots and dairy farms in the northwestern USA. *Prev Vet Med.* 1998 Apr 16;35(1):11-9. doi: 10.1016/s0167-5877(98)00050-6. PMID 9638776.
 47. Ahn CK, Russo, AJ, Howell KR, Holt NJ, Sellenriek PL, Rothbaum RJ, Beck AM, Luebbering LK, Tarr PL. Deer sausage: a newly identified vehicle of transmission of *Escherichia coli* O157:H7. *J Pediatr.* 2009 Oct;155(4):587-9. doi: 10.1016/j.jpeds.2009.02.051. PMID 19773004.
 48. Sánchez, Martínez R, García A, Vidal D, Blanco J, Blanco M, Blanco JE, Mora A, Herrera-León S, Echeita A, Alonso JM, Rey J. Detection and characterisation of O157:H7 and non-O157 Shiga toxin-producing *Escherichia coli* in wild boars. *Vet Microbiol.* 2010 Jul 14;143(2-4):420-3. doi: 10.1016/j.vetmic.2009.11.016. PMID 20005055.
 49. Renter DG, Sargent JM, Oberst RD, Samadpour M. Diversity, frequency, and persistence of *Escherichia coli* O157 strains from range cattle environments. *Appl Environ Microbiol.* 2003 Jan;69(1):542-7. doi: 10.1128/AEM.69.1.542-547.2003. PMID 12514039.
 50. Wallace JS, Cheasty T, Jones K. Isolation of vero cytotoxin-producing *Escherichia coli* O157 from wild birds. *J Appl Microbiol.* 1997 Mar;82(3):399-404. doi: 10.1046/j.1365-2672.1997.00378.x. PMID 12455905.
 51. Johnsen G, Wasteson Y, Heir E, Berget OI, Herikstad H. *Escherichia coli* O157:H7 in faeces from cattle, sheep and pigs in the southwest part of Norway during 1998 and 1999. *Int J Food Microbiol.* 2001 May 10;65(3):193-200. doi: 10.1016/s0168-1605(00)00518-3. PMID 11393688.
 52. Chase-Topping ME, McKendrick IJ, Pearce MC, MacDonald P, Matthews L, Halliday J, Allison L, Fenlon D, Low C, Gunn G, Woolhouse MEJ. Risk factors for the presence of high-level shedders of *Escherichia coli* O157 on Scottish farms. *J Clin Microbiol.* 2007 May;45(5):1594-603. doi: 10.1128/JCM.01690-06. PMID 17360845.
 53. Chase-Topping M, Gally D, Low C, Matthews L, Woolhouse M. Super-shedding and the link between human infection and livestock carriage of *Escherichia coli* O157. *Nat Rev Microbiol.* 2008;6(12):904–12. doi: 10.1038/nrmicro2029. PMID 19008890.
 54. Omisakin F, MacRae M, Ogden ID, Strachan NJC. Concentration and prevalence of *Escherichia coli* O157 in cattle feces at slaughter. *Appl Environ Microbiol.* 2003 May;69(5):2444-7. doi: 10.1128/AEM.69.5.2444-2447.2003. PMID 12732509.
 55. Katani R, Cote R, Kudva IT, DebRoy C, Arthur TM, Kapur V. Comparative genomics of two super-shedder isolates of *Escherichia coli* O157:H7. *PLoS One.* 2017 Aug 10;12(8):e0182940. doi: 10.1371/journal.pone.0182940. PMID 28797098.
 56. Butt S, Smith-Palmer A, Shand A, McDonald E, Allison L, Maund J, Fernandes A, Vistram B, Greig DR, Jenkins C, Elson R, Outbreak Control Team. Evidence of on-going transmission of Shiga toxin-producing *Escherichia coli* O157:H7 following a foodborne outbreak. *Epidemiol Infect.* 2021 Jun 7;149:e147. doi: 10.1017/S0950268821001278. PMID 34096488.
 57. Jenkins C, Bird PK, Wensley A, Wilkinson K, Aird H, Mackintosh A, Greig DR, Simpson A, Byrne L, Wilkinson R, Godbole G, Arunachalam N, Hughes GJ, Incident Management Team. Outbreak of STEC O157:H7 linked to a milk pasteurisation failure at a dairy farm in England, 2019. *Epidemiol Infect.* 2022 May 18;150:e114. doi: 10.1017/S0950268822000929. PMID 35581924.
 58. Coia JE, Jonston Y, Steers NG, Hanson MF. A survey of the prevalence of *Escherichia coli* O157 in raw meats, raw cow's milk and raw-milk cheeses in south-east Scotland. *Int J Food Microbiol.* 2001 May 21;66(1-2):63-9. doi: 10.1016/s0168-1605(00)00490-6. PMID 11407549.
 59. Treacy J, Jenkins C, Paranthaman K, Jorgensen F, Mueller-Doblies D, Anjum M, Kaindama L, Hartman H, Kirchner M, Carson T, Kar-Purkayastha I. Outbreak of Shiga toxin-producing *Escherichia coli* O157:H7 linked to raw drinking milk resolved by rapid application of

- advanced pathogen characterisation methods, England, August to October 2017. *Euro Surveill.* 2019;24(16):1800191. doi: 10.2807/1560-7917.ES.2019.24.16.1800191. PMID 31014418.
60. Kaindama L, Jenkins C, Aird H, Jorgensen F, Stoker K, Byrne L. *Epidemiol Infect.* 2021 May 6;149:e124. doi: 10.1017/S0950268821001072. PMID 33955833.
 61. Rowell S, King C, Jenkins C, Dallman TJ, Decraene V, Lamden K, Howard A, Featherstone CA, Cleary P. An outbreak of Shiga toxin-producing *Escherichia coli* serogroup O157 linked to a lamb-feeding event. *Epidemiol Infect.* 2016;144(12):2494–500. doi: 10.1017/S0950268816001229. PMID 27297133.
 62. Paunio M, Pebody R, Keskimäki M, Kokki M, Ruutu P, Oinonen S, Vuotari V, Siitonen A, Lahti E, Leinikki P. Swimming-associated outbreak of *Escherichia coli* O157:H7. *Epidemiol Infect.* 1999 Feb;122(1):1-5. doi: 10.1017/s0950268898001927. PMID 10098778.
 63. Harrison S, Kinra S. Outbreak of *Escherichia coli* O157 associated with a busy bathing beach. *Commun Dis Public Health.* 2004 Mar;7(1):47-50. PMID 15137281.
 64. Ihekweazu C, Carroll K, Adak B, Smith G, Pritchard GC, Gillespie IA, Verlander NQ, Harvey-Vince L, Reacher M, Edeghere O, Sultan B, Cooper R, Morgan G, Kinross PTN, Boxall NS, Iversen A, Bickler G. Large outbreak of verocytotoxin-producing *Escherichia coli* O157 infection in visitors to a petting farm in South East England, 2009. *Epidemiol Infect.* 2012 Aug;140(8):1400-13. doi: 10.1017/S0950268811002111. PMID 22093751.
 65. Richardson HY, Nicholas G, Lane C, Lake IR, Hunter PR. Microbiological surveillance of private water supplies in England: the impact of environmental and climate factors on water quality. *Water Res.* 2009 May;43(8):2159-68. doi: 10.1016/j.watres.2009.02.035. PMID 19303126.
 66. Elson R, Grace K, Vivancos R, Jenkins C, Adak GK, O'Brien SJ, Lake IR. A spatial and temporal analysis of risk factors associated with sporadic Shiga toxin-producing *Escherichia coli* O157 infection in England between 2009 and 2015. *Epidemiol Infect.* 2018 Nov;146(15):1928-1939. doi: 10.1017/S095026881800256X. PMID 30205851.
 67. Kintz E, Byrne L, Jenkins C, McCarthy N, Vivancos R, Hunter P. Outbreaks of Shiga Toxin-Producing *Escherichia coli* Linked to Sprouted Seeds, Salad, and Leafy Greens: A Systematic Review. *J Food Prot.* 2019 Nov;82(11):1950-1958. doi: 10.4315/0362-028X.JFP-19-014. PMID 31638410.
 68. Launders N, Locking ME, Hanson M, Willshar G, Charlett A, Salmon R, Cowden J, Harker KS, Adak GK. A large Great Britain-wide outbreak of STEC O157 phage type 8 linked to handling of raw leeks and potatoes. *Epidemiol Infect.* 2016 Jan;144(1):171-81. doi: 10.1017/S0950268815001016. PMID 26041509.
 69. Afza M, Hawker J, Thurston H, Gunn K, Orendi J. An outbreak of *Escherichia coli* O157 gastroenteritis in a care home for the elderly. *Epidemiol Infect.* 2006 Dec;134(6):1276-81. doi: 10.1017/S0950268806006546. PMID 16740198.
 70. O'Brien SJ, Murdoch PS, Riley AH, King I, Barr M, Murdoch S, Greig A, Main R, Reilly WJ, Thomson-Carter FM. A foodborne outbreak of Vero cytotoxin-producing *Escherichia coli* O157:H-phage type 8 in hospital. *J Hosp Infect.* 2001 Nov;49(3):167-72. doi: 10.1053/jhin.2001.1076. PMID 11716633.
 71. Cheasty T, Robertson R, Chart H, Mannion P, Syed Q, Garvery R, Rowe B. The use of serodiagnosis in the retrospective investigation of a nursery outbreak associated with *Escherichia coli* O157:H7. *J Clin Pathol.* 1998 Jul;51(7):498-501. doi: 10.1136/jcp.51.7.498. PMID 9797724.
 72. Bayliss L, Carr R, Edeghere O, Knapper E, Nye K, Harvey G, Adak G, Duggal H. School outbreak of *Escherichia coli* O157 with high levels of transmission, Staffordshire, England, February 2012. *J Public Health (Oxf).* 2016 Sep;38(3):e247-e253. doi: 10.1093/pubmed/fdv122. PMID 26364319.

73. Perna NT, Plunkett 3rd G, Burland V, Mau B, Glasner JD, Rose DJ, Mayhew GF, Evans PS, Gregor J, Kirkpatrick HA, Pósfai G, Hackett J, Klink S, Boutin A, Shao Y, Moller L, Grotbeck EJ, Davis NW, Lim A, Dimalanta ET, Potasmousis KD, Apodaca J, Anantharaman TS, Lin J, Yen G, Schwartz DC, Welch RA, Blattnew FR. Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. *Nature*. 2001 Jan 25;409(6819):529-33. doi: 10.1038/35054089. PMID 11206551.
74. Blattner FR, Plunkett G, Bloch CA, Perna NT, Burland V, Riley M, Collado-Vides J, Glasner JD, Rode CK, Mayhew GF, Gregor J, Davis NW, Kirkpatrick HA, Goeden MA, Rose DJ, Mau B, Shao Y. The complete genome sequence of *Escherichia coli* K-12. *Science*. 1997;277(5331):1453–62. doi: 10.1126/science.277.5331.1453. PMID 9278503.
75. Hayashi T, Makino K, Ohnishi M, Kurokawa K, Ishii K, Yokoyama K, Han, CG, Ohtsubo E, Nakayama K, Murata T, Tanaka M, Tobe T, Iida T, Takami H, Honda T, Sasakawa C, Ogasawara N, Yasunaga T, Kuhara S, Shiba T, Hattori M, Shingawa H. Complete genome sequence of enterohemorrhagic *Escherichia coli* O157:H7 and genomic comparison with a laboratory strain K-12. *DNA Res*. 2001;8(1):11–22. doi: 10.1093/dnares/8.1.11. PMID 11258796.
76. Feng P, Lampel KA, Karch H, Whittam TS. Genotypic and phenotypic changes in the emergence of *Escherichia coli* O157:H7. *J Infect Dis*. 1998;177(6):1750–3. doi: 10.1086/517438. PMID 9607864.
77. Feng PC, Monday SR, Lacher DW, Allison L, Siitonen A, Keys C, Eklund M, Nagano H, Karch H, Keen J, Whittam TS. Genetic diversity among clonal lineages within *Escherichia coli* O157:H7 stepwise evolutionary model. *Emerg Infect Dis*. 2007 Nov;13(11):1701-6. doi: 10.3201/eid1311.070381. PMID 18217554.
78. Leopold SR, Magrini VM, Holt NJ, Shaikh N, Mardis ER, Cagno J, Ogura Y, Iguchi AM, Hayash T, Mellmann A, Karch H, Besser TE, Sawyer SA, Whittam TS, Tarr PI. A precise reconstruction of the emergence and constrained radiations of *Escherichia coli* O157 portrayed by backbone concatenomic analysis. *Proc Natl Acad Sci USA*. 2009 May 26;106(21):8713-8. doi: 10.1073/pnas.0812949106. PMID 19439656.
79. Kyle JL, Cummings CA, Parker CT, Quiñones B, Vatta P, Newton E, Huynh S, Swimley M, Degoricija L, Barker M, Fontanoz S, Nguyen K, Patel R, Fang R, Tebbs R, Petrauskene O, Furtado M, Mandrell RE. *Escherichia coli* serotype O55:H7 diversity supports parallel acquisition of bacteriophage at Shiga toxin phage insertion sites during evolution of the O157:H7 lineage. *J Bacteriol*. 2012 Apr;194(8):1885-96. doi: 10.1128/JB.00120-12. PMID 22328665.
80. Dallman TJ, Ashton PM, Byrne L, Perry NT, Petrovska L, Ellis R, Allison L, Hanson M, Holmes A, Gunn GJ, Chase-Topping M, Woolhouse MEJ, Grant KA, Gally DL, Wain J, Jenkins C. 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(3):e000029. doi: 10.1099/mgen.0.000029. PMID 28348814.
81. Shaaban S, Cowley LA, McAteer SP, Jenkins C, Dallman TJ, Bono JL, Gally DL. Evolution of a zoonotic pathogen: investigating prophage diversity in enterohaemorrhagic *Escherichia coli* O157 by long-read sequencing. *Microb Genom*. 2016;2(12):e000096. doi: 10.1099/mgen.0.000096. PMID 28348836.
82. Yang Z, Kovar J, Kim J, Nietfeldt J, Smith DR, Moxley RA, Olson ME, Fey PD, Benson AK. Identification of common subpopulations of non-sorbitol-fermenting, beta-glucuronidase-negative *Escherichia coli* O157:H7 from bovine production environments and human clinical samples. *Appl Environ Microbiol*. 2004 Nov;70(11):6846-54. doi: 10.1128/AEM.70.11.6846-6854.2004. PMID 15528552.
83. Zadik PM, Chapman PA, Siddons CA. Use of tellurite for the selection of verocytotoxigenic *Escherichia coli* O157. *J Med Microbiol*. 1993 Aug;39(2):155-8. doi: 10.1099/00222615-39-2-155. PMID 8345511.

84. March SB, Ratnam S. Sorbitol-MacConkey medium for detection of *Escherichia coli* O157:H7 associated with hemorrhagic colitis. *J Clin Microbiol.* 1986 May;23(5):869-72. doi: 10.1128/jcm.23.5.869-872.1986. PMID 3519658.
85. Bettelheim KA, Thompson CJ. New method of serotyping *Escherichia coli*: implementation and verification. *J Clin Microbiol.* 1987 May;25(5):781-6. doi: 10.1128/jcm.25.5.781-786.1987. PMID 2438301.
86. Orskov I, Orskov F, Jann B, Jann K. Serology, chemistry, and genetics of O and K antigens of *Escherichia coli*. *Bacteriol Rev.* 1977 Sep;41(3):667-710. doi: 10.1128/br.41.3.667-710.1977. PMID 334154.
87. Perelle S, Dilasser F, Grout J, Fach P. Detection by 5'-nuclease PCR of Shiga-toxin producing *Escherichia coli* O26, O55, O91, O103, O111, O113, O145 and O157:H7, associated with the world's most frequent clinical cases. *Mol Cell Probes.* 2004 Jun;18(3):185-92. doi: 10.1016/j.mcp.2003.12.004. PMID 15135453.
88. Ahmed R, Bopp C, Borczyk A, Kasatiya S. Phage-typing scheme for *Escherichia coli* O157:H7. *J Infect Dis.* 1987;155(4):806-9. doi: 10.1093/infdis/155.4.806. PMID 3546523.
89. Khakhria R, Duck D, Lior H. Extended phage-typing scheme for *Escherichia coli* O157:H7. *Epidemiol Infect.* 1990;105(3):511-20. doi: 10.1017/s0950268800048135. PMID 2249715.
90. Böhm H, Karch H. DNA fingerprinting of *Escherichia coli* O157:H7 strains by pulsed-field gel electrophoresis. *J Clin Microbiol.* 1992 Aug;30(8):2169-72. doi: 10.1128/jcm.30.8.2169-2172.1992. PMID 1500528.
91. Dallman TJ, Byrne L, Ashton PM, Cowley LA, Perry NT, Adak G, Petrovska L, Ellis RJ, Elson R, Underwood A, Green J, Hanage WP, Jenkins C, Grant K, Wain J. Whole-genome sequencing for national surveillance of Shiga toxin-producing *Escherichia coli* O157. *Clin Infect Dis.* 2015 Aug 1;61(3):305-12. doi: 10.1093/cid/civ318. PMID 25888672.
92. Maiden MC, Bygraves Ja, Feil E, Morelli G, Russel JE, Urwin R, Zhang Q, Zhou J, Zurth K, Caugant DA, Feavers IM, Achtman M, Spratt BG. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc Natl Acad Sci USA.* 1998 Mar 17;95(6):3140-5. doi: 10.1073/pnas.95.6.3140. PMID 9501229.
93. Ashton PM, Perry N, Ellis R, Petrovska L, Wain J, Grant KA, Jenkins C, Dallman TJ. Insight into Shiga toxin genes encoded by *Escherichia coli* O157 from whole genome sequencing. *PeerJ.* 2015;3:e739. doi: 10.7717/peerj.739. PMID 25737808.
94. Dallman T, Ashton P, Schafer U, Jironkin A, Painset A, Shaaban S, Hartman H, Myers R, Underwood A, Jenkins C, Grant K. SnapperDB: a database solution for routine sequencing analysis of bacterial isolates. *Bioinformatics.* 2018 Sep 1;34(17):3028-3029. doi: 10.1093/bioinformatics/bty212. PMID 29659710.
95. Whittam TS, Wilson RA. Genetic relationships among pathogenic *Escherichia coli* of serogroup O157. *Infect Immun.* 1988 Sep;56(9):2467-73. doi: 10.1128/iai.56.9.2467-2473.1988. PMID 2457555.
96. Manning SD, Motiwala AF, Springman AC, Qi W, Lacher DW, Ouellette LM, Mladonickly JM, Somsel P, Rudrik JT, Dietrich SE, Xhang W, Swaminathan B, Alland D, Whittam TS. Variation in virulence among clades of *Escherichia coli* O157:H7 associated with disease outbreaks. *Proc Natl Acad Sci USA.* 2008 Mar 25;105(12):4868-73. doi: 10.1073/pnas.0710834105. PMID 18332430.
97. Yokoyama E, Hirai S, Hashimoto R, Uchimura M. Clade analysis of enterohemorrhagic *Escherichia coli* serotype O157:H7/H- strains and hierarchy of their phylogenetic relationships. *Infect Genet Evol.* 2012 Dec;12(8):1724-8. doi: 10.1016/j.meegid.2012.07.003. PMID 22846398.
98. Dallman TJ, Greig DR, Gharbia SE, Jenkins C. Phylogenetic structure of Shiga toxin-producing *Escherichia coli* O157:H7 from sub-lineage to SNPs. *Microb Genom.* 2021 Mar;7(3):mgen000544. doi: 10.1099/mgen.0.000544. PMID 33720818.

99. Dallman TJ, Jalava K, Verlander NQ, Gally D, Jenkins C, Godbole G, Gharbia S. Identification of domestic reservoirs and common exposures in an emerging lineage of Shiga toxin-producing *Escherichia coli* O157:H7 in England: a genomic epidemiological analysis. *Lancet Microbe*. 2022 Aug;3(8):e606-e615. doi: 10.1016/S2666-5247(22)00089-1. PMID 35760076.
100. Adams NL, Byrne L, Smith GA, Elson R, Harris JP, Salmon R, Smith R, O'Brien SJ, Adak GK, Jenkins C. Shiga Toxin-Producing *Escherichia coli* O157, England and Wales, 1983-2012. *Emerg Infect Dis*. 2016 Apr;22(4):590-7. doi: 10.3201/eid2204.151485. PMID 26982243.
101. Byrne L, Jenkins C, Launders N, Elson R, Adak GK. The epidemiology, microbiology and clinical impact of Shiga toxin-producing *Escherichia coli* in England, 2009-2012. *Epidemiol Infect*. 2015 Dec;143(16):3475-87. doi: 10.1017/S0950268815000746. PMID 25920912.
102. Franz E, Rotariu O, Lopes BS, MacRae M, Bono JL, Laing C, Gannon V, Söderlund R, van Hoek AHAM, Friesema I, French NP, George T, Biggs PJ, Jaros P, Rivas M, Chinen I, Campos J, Jernberg C, Gobius K, Mellor GE, Chandry PS, Perez-Reche P, Forbes KJ, Strachan NJC. Phylogeographic Analysis Reveals Multiple International transmission Events Have Driven the Global Emergence of *Escherichia coli* O157:H7. *Clin Infect Dis*. 2019 Jul 18;69(3):428-437. doi: 10.1093/cid/ciy919. PMID 30371758.
103. Byrne L, Dallman TJ, Adams N, Mikhail AFW, McCarthy N, Jenkins C. Highly Pathogenic Clone of Shiga Toxin-Producing *Escherichia coli* O157:H7, England and Wales. *Emerg Infect Dis*. 2018 Dec;24(12):2303-2308. doi: 10.3201/eid2412.180409. PMID 30457532.
104. 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 Jan;24(4):1800346. doi: 10.2807/1560-7917.ES.2019.24.4.1800346. PMID 30696532.
105. Goh S, Newman C, Knowles M, Bolton FJ, Hollyoak V, Richards S, Daley P, Counter D, Smith HR, Keppie N. *E. coli* O157 phage type 21/28 outbreak in North Cumbria associated with pasteurized milk. *Epidemiol Infect*. 2002 Dec;129(3):451-7. doi: 10.1017/s0950268802007835. PMID 12558327.
106. Wilson D, Dolan G, Aird H, Sorrell S, Dallman TJ, Jenkins C, Robertson L, Gorton R. Farm-to-fork investigation of an outbreak of Shiga toxin-producing *Escherichia coli* O157. *Microb Genom*. 2018 Mar;4(3):e000160. doi: 10.1099/mgen.0.000160. PMID 29488865.
107. Gobin M, Hawker J, Cleary P, Inns T, Gardiner D, Mikhail A, McCormick J, Elson R, Ready D, Dallman T, Roddick I, Hall I, Willis C, Crook P, Godbole G, Turbin-Delic D, Oliver I. National outbreak of Shiga toxin-producing *Escherichia coli* O157:H7 linked to mixed salad leaves, United Kingdom, 2016. *Euro Surveill*. 2018 May;23(18):17-00197. doi: 10.2807/1560-7917.ES.2018.23.18.17-00197. PMID 29741151.
108. Campbell A. The future of bacteriophage biology. *Nat Rev Genet*. 2003 Jun;4(6):471-7. doi: 10.1038/nrg1089. PMID 12776216.
109. Ramisetty BCM, Sudhakari PA. Bacterial 'Grounded' Prophages: Hotspots for Genetic Renovation and Innovation. *Front Genet*. 2019;10:65. doi: 10.3389/fgene.2019.00065. PMID 30809245.
110. Bobay LM, Rocha EPC, Touchon M. The adaptation of temperate bacteriophages to their host genomes. *Mol Biol Evol*. 2013 Apr;30(4):737-51. doi: 10.1093/molbev/mss279. PMID 23243039.
111. Asadulghani M, Ogura Y, Ooka T, Itoh T, Sawaguchi A, Iguchi A, Nakayama K, Hayashi T. The defective prophage pool of *Escherichia coli* O157: prophage-prophage interactions potentiate horizontal transfer of virulence determinants. *PLoS Pathog*. 2009;5(5):e1000408. doi: 10.1371/journal.ppat.1000408. PMID 19412337.
112. Koch C, Hertwig S, Appel B. Nucleotide sequence of the integration site of the temperate bacteriophage 6220, which carries the Shiga toxin gene *stx*_(10x3). *J Bacteriol*. 2003;185(21):6463-6. doi: 10.1128/JB.185.21.6463-6466.2003. PMID 14563882.

113. Creuzburg K, Köhler B, Hempel H, Schreier P, Jacobs E, Schmidt H. Genetic structure and chromosomal integration site of the cryptic prophage CP-1639 encoding Shiga toxin 1. *Microbiology*. 2005;151(Pt 3):941–50. doi: 10.1099/mic.0.27632-0. PMID 15758239.
114. Besser TE, Shaikh N, Holt NJ, Tarr PI, Konkel ME, Malik-Kale P, Walsh CW, Whittam TS, Bono JL. Greater diversity of Shiga toxin-encoding bacteriophage insertion sites among *Escherichia coli* O157:H7 isolates from cattle than in those from humans. *Appl Environ Microbiol*. 2007;73(3):671–9. doi: 10.1128/AEM.01035-06. PMID 17142358.
115. Ogura Y, Ooka T, Iguchi A, Toh H, Asadulghani M, Oshima K, Kodama T, Abe H, Nakayama K, Kurokawa K, Tobe T, Hattori M, Hayashi T. Comparative genomics reveal the mechanism of the parallel evolution of O157 and non-O157 enterohemorrhagic *Escherichia coli*. *Proc Natl Acad Sci USA*. 2009;106(42):17939–44. doi: 10.1073/pnas.0903585106. PMID 19815525.
116. Steyert SR, Sahl JW, Fraser CM, Teel LD, Scheutz F, Rasko DA. Comparative genomics and *stx* phage characterization of LEE-negative Shiga toxin-producing *Escherichia coli*. *Front Cell Infect Microbiol*. 2012;2:133. doi: 10.3389/fcimb.2012.00133. PMID 23162798.
117. Ogura Y, Mondal SI, Islam MR, Mako T, Arisawa K, Katsura K, Ooka T, Gotoh Y, Murase K, Ohinishi M, Hayashi T. The Shiga toxin 2 production level in enterohemorrhagic *Escherichia coli* O157:H7 is correlated with the subtypes of toxin-encoding phage. *Sci Rep*. 2015;16;5:16663. doi: 10.1038/srep16663. PMID 26567959.
118. Strockbine NA, Narques LR, Newland JW, Smith HW, Holmes RK, O'Brien AD. Two toxin-converting phages from *Escherichia coli* O157:H7 strain 933 encode antigenically distinct toxins with similar biologic activities. *Infect Immun*. 1986 Jul;53(1):135-40. doi: 10.1128/iai.53.1.135-140.1986. PMID 3522426.
119. Johnansen BK, Wasteson Y, Granum PE, Brynestad S. Mosaic structure of Shiga-toxin-2-encoding phages isolated from *Escherichia coli* O157:H7 indicates frequent gene exchange between lambdoid phage genomes. *Microbiology*. 2001 Jul;147(Pt 7):1929-1936. doi: 10.1099/00221287-147-7-1929. PMID 11429469.
120. Schmidt H. Shiga-toxin-converting bacteriophages. *Res Microbiol*. 2001;152(8):687–95. doi: 10.1016/s0923-2508(01)01249-9. PMID 11686382.
121. Waldor MK, Friedman DI. Phage regulatory circuits and virulence gene expression. *Curr Opin Microbiol*. 2005;8(4):459–65. doi: 10.1016/j.mib.2005.06.001. PMID 15979389.
122. Wong CS, Jelacic S, Habeeb RL, Watkins SL, Tarr PI. The risk of the hemolytic-uremic syndrome after antibiotic treatment of *Escherichia coli* O157:H7 infections. *N Engl J Med*. 2000;342(26):1930–6. doi: 10.1056/NEJM200006293422601. PMID 10874060.
123. Carter AO, Nprczyk AA, Carlson Ja, Harvery B, Hockin JC, Karmali MA, Krishnan C, Korn DA, Lior H. A severe outbreak of *Escherichia coli* O157:H7--associated hemorrhagic colitis in a nursing home. *N Engl J Med*. 1987 Dec 10;317(24):1496-500. doi: 10.1056/NEJM198712103172403. PMID 3317047.
124. Iguchi A, Iyoda S, Terajima J, Watanabe H, Osawa R. Spontaneous recombination between homologous prophage regions causes large-scale inversions within the *Escherichia coli* O157:H7 chromosome. *Gene*. 2006;372:199–207. doi: 10.1016/j.gene.2006.01.005. PMID 16516407.
125. Raeside C, Gaffé J, Deatherage DE, Tenaillon O, Briska AM, Ptashkin RN, Cruveiller S, Médigue C, Lenski RE, Barrick JE, Schneider D. Large chromosomal rearrangements during a long-term evolution experiment with *Escherichia coli*. *MBio*. 2014;5(5):e01377-01314. doi: 10.1128/mBio.01377-14. PMID 25205090.
126. Lee H, Doak TG, Popodi E, Foster PL, Tang H. Insertion sequence-caused large-scale rearrangements in the genome of *Escherichia coli*. *Nucleic Acids Res*. 2016;44(15):7109–19. doi: 10.1093/nar/gkw647. PMID 27431326.
127. Eppinger M, Mammel MK, Leclerc JE, Ravel J, Cebula TA. Genomic anatomy of *Escherichia coli* O157:H7 outbreaks. *Proc Natl Acad Sci USA*. 2011 Dec 13;108(50):20142-7. doi: 10.1073/pnas.1107176108. PMID 22135463.

128. Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, Fiddes CA, Hutchison CA, Slocombe PM, Smith M. Nucleotide sequence of bacteriophage phi X174 DNA. *Nature*. 1977 Feb 24;265(5596):687-95. doi: 10.1038/265687a0. PMID 870828.
129. Ju J, Kim DH, Bi L, Meng Q, Bai X, Li Z, Li X, Marma MS, Shi S, Wu J, Edwards JR, Romu Am Turro NJ. Four-color DNA sequencing by synthesis using cleavable fluorescent nucleotide reversible terminators. *Proc Natl Acad Sci USA*. 2006 Dec 26;103(52):19635-40. doi: 10.1073/pnas.0609513103. PMID 17170132.
130. 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 Feb 22;56(3):e01388-17. doi: 10.1128/JCM.01388-17. PMID 29263202.
131. Lupolova N, Dallman TJ, Matthews L, Bono JL, Gally DL. Support vector machine applied to predict the zoonotic potential of *E. coli* O157 cattle isolates. *Proc Natl Acad Sci USA*. 2016 Oct 4;113(40):11312-11317. doi: 10.1073/pnas.1606567113.
132. Kasianowicz JJ, Brandin E, Branton D, Deamer DW. Characterization of individual polynucleotide molecules using a membrane channel. *Proc Natl Acad Sci USA*. 1996 Nov 26;93(24):13770-3. doi: 10.1073/pnas.93.24.13770. PMID 8943010.
133. Simpson JT, Workman RE, Zuzarte PC, David M, Dursi LJ, Timp W. Detecting DNA cytosine methylation using nanopore sequencing. *Nat Methods*. 2017 Apr;14(4):407-410. doi: 10.1038/nmeth.4184. PMID 28218898.
134. Besser J, Carleton HA, Gerner-Smidt P, Lindsey RL, Trees E. Next-generation sequencing technologies and their application to the study and control of bacterial infections. *Clin Microbiol Infect*. 2018;24(4):335–41. doi: 10.1016/j.cmi.2017.10.013. PMID 29074157.
135. Duan M, Hunter SS, Minnich SA, Fagnan MW, New DD, Hovde CJ, Sheng H. Complete Genome Sequence of Broad-Host-Range Shiga Toxin-Converting Bacteriophage SH2026Stx1, Isolated from *Escherichia coli* O157:H7. *Genome Announc*. 2018;6(25):e00490-18. doi: 10.1128/genomeA.00490-18. PMID 29930034.
136. Peritz A, Paoli GC, Chen C-Y, Gehring AG. Serogroup-level resolution of the “Super-7” Shiga toxin-producing *Escherichia coli* using nanopore single-molecule DNA sequencing. *Anal Bioanal Chem*. 2018;410(22):5439–44. doi: 10.1007/s00216-018-0877-1. PMID 29374775.
137. Teng L, Ginn A, Jeon S, Kang M, Jeong KC. Complete Genome Sequence of an *Escherichia coli* O157:H7 Strain Isolated from a Super-Shedder Steer. *Genome Announc*. 2016;4(2):e00258-16. doi: 10.1128/genomeA.00258-16. PMID 27056233.
138. Gonzalez-Escalona N, Allard MA, Brown EW, Sharma S, Hoffmann M. Nanopore sequencing for fast determination of plasmids, phages, virulence markers, and antimicrobial resistance genes in Shiga toxin-producing *Escherichia coli*. *PLoS One*. 2019;14(7):e0220494. doi: 10.1371/journal.pone.0220494. PMID 31361781.
139. Cowley LA, Dallman TJ, Fitzgerald S, Irvine N, Rooney PJ, McAteer SP, Day M, Perry NT, Bono JL, Jenkins C, Gally DL. Short-term evolution of Shiga toxin-producing *Escherichia coli* O157:H7 between two food-borne outbreaks. *Microb Genom*. 2016;2(9):e000084. doi: 10.1099/mgen.0.000084. PMID 28348875.
140. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*. 2014 Aug 1;30(15):2114-20. doi: 10.1093/bioinformatics/btu170. PMID 24695404.
141. De Coster W, D’Hert S, Schultz DT, Cruts M, Broeckhoven CV. NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics*. 2018 Aug 1;34(15):2666-2669. doi: 10.1093/bioinformatics/bty149. PMID 29547981.
142. Wick RR, <https://github.com/rrwick/Porechop>
143. Wick RR, <https://github.com/rrwick/Filtlong>

144. Kolmogorov M, Yuan J, Lin Y, Pevzner PA. Assembly of long, error-prone reads using repeat graphs. *Nat Biotechnol.* 2019 May;37(5):540-546. doi: 10.1038/s41587-019-0072-8. PMID 30936562.
145. Loman NJ, Quick J, Simpson JT. A complete bacterial genome assembled *de novo* using only nanopore sequencing data. *Nat Methods.* 2015 Aug;12(8):733-5. doi: 10.1038/nmeth.3444. PMID 26076426.
146. Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics.* 2018 Sep 15;34(18):3094-3100. doi: 10.1093/bioinformatics/bty191. PMID 29750242.
147. Li H, Handsaker B, Wysoker A, Fennel T, Ruan J, Homer N, Marth G, Abecasis G, Durban R, 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics.* 2009 Aug 15;25(16):2078-9. doi: 10.1093/bioinformatics/btp352. PMID 19505943.
148. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, Earl AM. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One.* 2014 Nov 19;9(11):e112963. doi: 10.1371/journal.pone.0112963. PMID 25409509.
149. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics.* 2009 Jul 15;25(14):1754-60. doi: 10.1093/bioinformatics/btp324. PMID 19451168.
150. Vaser R, Sović I, Nagarajan N, Šikić M. Fast and accurate *de novo* genome assembly from long uncorrected reads. *Genome Res.* 2017 May;27(5):737-746. doi: 10.1101/gr.214270.116. PMID 28100585.
151. Oxford Nanopore Technologies, <https://github.com/nanoporetech/medaka>
152. Oxford Nanopore Technologies, <https://github.com/nanoporetech/pomoxis>
153. Oxford Nanopore Technologies, <https://github.com/nanoporetech/medaka>
154. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin Vm, Nikolenko SI, Pham S, Prjibelsji AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol.* 2012 May;19(5):455-77. doi: 10.1089/cmb.2012.0021. PMID 22506599.
155. Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics.* 2014 Jul 15;30(14):2068-9. doi: 10.1093/bioinformatics/btu153. PMID 24642063.
156. Tatusova T, DiCuccio M, Badretdin A, Chetvernin V, Nawrocki EP, Zaslavsky L, Lomsadze A, Pruitt KD, Borodovsky M, Ostell J. NCBI prokaryotic genome annotation pipeline. *Nucleic Acids Res.* 2016 Aug 19;44(14):6614-24. doi: 10.1093/nar/gkw569. PMID 27342282.
157. Schwengers O, Jelonek L, Dieckmann MA, Beyvers S, Blom J, Goesmann A. Bakta: rapid and standardized annotation of bacterial genomes via alignment-free sequence identification. *Microb Genom.* 2021 Nov;7(11):000685. doi: 10.1099/mgen.0.000685. PMID 34739369.
158. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. Circos: an information aesthetic for comparative genomics. *Genome Res.* 2009 Sep;19(9):1639-45. doi: 10.1101/gr.092759.109. PMID 19541911.
159. UKHSA, <https://github.com/ukhsa-collaboration/kmerid>
160. Tewolde R, Dallman T, Schaefer U, Sheppard CL, Ashton P, Pichon B, Ellington M, Swift C, Green J, Underwood A. MOST: a modified MLST typing tool based on short read sequencing. *PeerJ.* 2016 Aug 17;4:e2308. doi: 10.7717/peerj.2308. PMID 27602279.
161. UKHSA, https://github.com/ukhsa-collaboration/gene_finder
162. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods.* 2012 Mar 4;9(4):357-9. doi: 10.1038/nmeth.1923. PMID 22388286.

163. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 2010 Sep;20(9):1297-303. doi: 10.1101/gr.107524.110. PMID 20644199.
164. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* 2014 Mar 3;15(3):R46. doi: 10.1186/gb-2014-15-3-r46. PMID 24580807.
165. Page AJ, Keane AJ. Rapid multi-locus sequence typing direct from uncorrected long reads using Krocus. *PeerJ.* 2018 Jul 31;6:e5233. doi: 10.7717/peerj.5233. PMID 30083440.
166. Milne I, Bayer M, Cardle L, Shaw P, Stephen G, Wright F, Marshall D. Tablet--next generation sequence assembly visualization. *Bioinformatics.* 2010 Feb 1;26(3):401-2. doi: 10.1093/bioinformatics/btp666. PMID 19965881.
167. Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA, Bentley SD, Parkhill J, Harris SR. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Res.* 2015 Feb 18;43(3):e15. doi: 10.1093/nar/gku1196. PMID 25414349.
168. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics.* 2014 May 1;30(9):1312-3. doi: 10.1093/bioinformatics/btu033. PMID 24451623.
169. Wick RR, Judd LM, Gorrie CL, Holt KE. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Comput Biol.* 2017 Jun 8;13(6):e1005595. doi: 10.1371/journal.pcbi.1005595. PMID 28594827.
170. Yara DA, Greig DR, Gally DL, Dallman TJ, Jenkins C. Comparison of Shiga toxin-encoding bacteriophages in highly pathogenic strains of Shiga toxin-producing *Escherichia coli* O157:H7 in the UK. *Microb Genom.* 2020 Mar;6(3):e000334. doi: 10.1099/mgen.0.000334. PMID 32100710.
171. Greig DR, Jenkins C, Gharbia S, Dallman TJ. Comparison of single-nucleotide variants identified by Illumina and Oxford Nanopore technologies in the context of a potential outbreak of Shiga toxin-producing *Escherichia coli*. *Gigascience.* 2019 Aug 1;8(8):giz104. doi: 10.1093/gigascience/giz104. PMID 31433830.
172. Mikheyev AS, Tin MMY. A first look at the Oxford Nanopore MinION sequencer. *Mol Ecol Resour.* 2014 Nov;14(6):1097-102. doi: 10.1111/1755-0998.12324. PMID 25187008.
173. Laver T, Harrison J, O'Neill PA, Moore K, Farbos A, PAszkiewicz K, Studholme DJ. Assessing the performance of the Oxford Nanopore Technologies MinION. *Biomol Detect Quantif.* 2015 Mar;3:1-8. doi: 10.1016/j.bdq.2015.02.001. PMID 26753127.
174. Magi A, Semeraro R, Mingrino A, Giusti B, D'Aurizio R. Nanopore sequencing data analysis: state of the art, applications and challenges. *Brief Bioinform.* 2018 Nov 27;19(6):1256-1272. doi: 10.1093/bib/bbx062. PMID 28637243.
175. Rang FJ, Kloosterman WP, Ridder J. From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. *Genome Biol.* 2018 Jul 13;19(1):90. doi: 10.1186/s13059-018-1462-9. PMID 30005597.
176. Cali DS, Kim JS, Ghose S, Alkan C, Mutlu O. Nanopore sequencing technology and tools for genome assembly: computational analysis of the current state, bottlenecks and future directions. *Brief Bioinform.* 2019 Jul 19;20(4):1542-1559. doi: 10.1093/bib/bby017. PMID 29617724.
177. Marinus MG. DNA methylation in *Escherichia coli*. *Annu Rev Genet.* 1987;21:113-31. doi: 10.1146/annurev.ge.21.120187.000553. PMID 3327459.
178. Gomez-Eichelmann MC, Levy-Mustri A, Ramirez-Santos J. Presence of 5-methylcytosine in CC(A/T)GG sequences (Dcm methylation) in DNAs from different bacteria. *J Bacteriol.* 1991 Dec;173(23):7692-4. doi: 10.1128/jb.173.23.7692-7694.1991.

179. Jenkins C, Dallman TJ, Launders N, Willis C, Byrne L, Jorgensen F, Eppinger M, Adak GK, Aird H, Elviss N, Grant KA, Morgan D, McLauchlin J. Public Health Investigation of Two Outbreaks of Shiga Toxin-Producing *Escherichia coli* O157 Associated with Consumption of Watercress. *Appl Environ Microbiol*. 2015 Jun 15;81(12):3946-52. doi: 10.1128/AEM.04188-14. PMID 25841005.
180. Uhlich GA, Sinclair JR, Warren NG, Chmielecki WA, Fratamico P. Characterization of Shiga toxin-producing *Escherichia coli* isolates associated with two multistate food-borne outbreaks that occurred in 2006. *Appl Environ Microbiol*. 2008 Feb;74(4):1268-72. doi: 10.1128/AEM.01618-07. PMID 18083883.
181. Michino H, Araki K, Minami S, Takaya S, Sakai N, Miyazaki M, Ono A, Yanagawa H. Massive outbreak of *Escherichia coli* O157:H7 infection in schoolchildren in Sakai City, Japan, associated with consumption of white radish sprouts. *Am J Epidemiol*. 1999 Oct 15;150(8):787-96. doi: 10.1093/oxfordjournals.aje.a010082. PMID 10522649.
182. Wick RR, Judd ML, Holt KE. Deepbiner: Demultiplexing barcoded Oxford Nanopore reads with deep convolutional neural networks. *PLoS Comput Biol*. 2018 Nov 20;14(11):e1006583. doi: 10.1371/journal.pcbi.1006583. PMID 30458005.
183. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NG, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res*. 2017 May;27(5):722-736. doi: 10.1101/gr.215087.116. PMID 28298431.
184. Hunt M, De Silva N, Otto TD, Parkhill J, Keane JA, Harris SR. Circlator: automated circularization of genome assemblies using long sequencing reads. *Genome Biol*. 2015 Dec 29;16:294. doi: 10.1186/s13059-015-0849-0. PMID 26714481.
185. Arndt D, Grant JR, Marcu A, Sajed T, Pon A, Liang Y, Wishart DS. PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic Acids Res*. 2016 Jul 8;44(W1):W16-21. doi: 10.1093/nar/gkw387. PMID 27141966.
186. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, Phillippy AM. Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol*. 2016 Jun 20;17(1):132. doi: 10.1186/s13059-016-0997-x. PMID 27323842.
187. Andrew Rambaut, <https://github.com/rambaut/figtree>
188. Sullivan M, Petty NK, Beatson SA. Easyfig: a genome comparison visualizer. *Bioinformatics*. 2011 Apr 1;27(7):1009-10. doi: 10.1093/bioinformatics/btr039. PMID 21278367.
189. Schutz K, Cowley LA, Shaaban S, Carroll A, McNamara E, Gally DL, Godbole G, Jenkins C, Dallman TJ. Evolutionary Context of Non-Sorbitol-Fermenting Shiga Toxin-Producing *Escherichia coli* O55:H7. *Emerg Infect Dis*. 2017 Dec;23(12):1966-1973. doi: 10.3201/eid2312.170628. PMID 29148397.
190. Mikhail AFW, Jenkins C, Dallman TJ, Inns T, Douglas A, Martín AIC, Fox A, Cleary P, Elson R, Hawker J. 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 Jan;146(2):187-196. doi: 10.1017/S0950268817002874. PMID 29248018.
191. Sharp A, Smout E, Byrne L, Greenwood R, Abdoolah R, Hutchinshon C, Jenkins C, Arunachalam N, Padfield S, Hughes G, Gent M. An outbreak of Shiga toxin-producing *Escherichia coli* O157:H7 linked to a mud-based obstacle course, England, August 2018. *Zoonoses Public Health*. 2020 Aug;67(5):467-473. doi: 10.1111/zph.12744. PMID 32564515.
192. Sereika M, Kirkengaard RH, Karst SM, Michaelsen TY, Sørensen EA, Wollenberg RD, Albertsen M. Oxford Nanopore R10.4 long-read sequencing enables the generation of near-finished bacterial genomes from pure cultures and metagenomes without short-read or reference polishing. *Nat Methods*. 2022 Jul;19(7):823-826. doi: 10.1038/s41592-022-01539-7. PMID 35789207.
193. Foster-Nyarko E, Cottingham H, Wick RR, Judd LM, Lam MMC, Wyres LK, Stanton TD, Tsang KK, David S, Anaensen DM, Brisse S, Holt KE. Nanopore-only assemblies for genomic

- surveillance of the global priority drug-resistant pathogen, *Klebsiella pneumoniae*. *Microb Genom*. 2023 Feb;9(2):mgen000936. doi: 10.1099/mgen.0.000936. PMID 36752781.
194. Wick RR, Judd LM, Holt KE. Assembling the perfect bacterial genome using Oxford Nanopore and Illumina sequencing. *PLoS Comput Biol*. 2023 Mar 2;19(3):e1010905. doi: 10.1371/journal.pcbi.1010905. PMID 36862631.
 195. Chattaway Ma, Dallman TJ, Gentle A, Wright MJ, Long SE, Ashton PM, Perry NT, Jenkins C. Whole Genome Sequencing for Public Health Surveillance of Shiga Toxin-Producing *Escherichia coli* Other than Serogroup O157. *Front Microbiol*. 2016 Mar 3;7:258. doi: 10.3389/fmicb.2016.00258. PMID 26973632.
 196. Greig DR, Jenkins C, Gharbia SE, Dallman TJ. Analysis of a small outbreak of Shiga toxin-producing *Escherichia coli* O157:H7 using long-read sequencing. *Microb Genom*. 2021 Mar;7(3):mgen000545. doi: 10.1099/mgen.0.000545. PMID 33683192.
 197. Carver T, Harris SR, Berriman M, Parkhill J, McQuillan JA. Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data. *Bioinformatics*. 2012 Feb 15;28(4):464-9. doi: 10.1093/bioinformatics/btr703. PMID 22199388.
 198. Carattoli A, Hasman H. PlasmidFinder and *In silico* pMLST: Identification and typing of plasmid replicons in whole-genome sequencing (WGS). *Methods Mol Biol*. 2020:2075:285-294. doi: 10.1007/978-1-4939-9877-7_20. PMID 31584170.
 199. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol*. 1990 Oct 5;215(3):403-10. doi: 10.1016/S0022-2836(05)80360-2. PMID 2231712.
 200. Wick RR, Judd LM, Holt KE. Performance of neural network basecalling tools for Oxford Nanopore sequencing. *Genome Biol*. 2019 Jun 24;20(1):129. doi: 10.1186/s13059-019-1727-y. PMID 31234903.
 201. Dohm JC, Peters P, Stralis-Pavese N, Himmelbauer H. Benchmarking of long-read correction methods. *NAR Genom Bioinform*. 2020 May 25;2(2):lqaa037. doi: 10.1093/nargab/lqaa037. PMID 33575591.
 202. Shah RN, Ruthenburg AJ. Sequence deeper without sequencing more: Bayesian resolution of ambiguously mapped reads. *PLoS Comput Biol*. 2021 Apr 19;17(4):e1008926. doi: 10.1371/journal.pcbi.1008926. PMID 33872311.
 203. Nakamura K, Ogura Y, Gotoh Y, Hayashi T. Prophages integrating into prophages: A mechanism to accumulate type III secretion effector genes and duplicate Shiga toxin-encoding prophages in *Escherichia coli*. *PLoS Pathog*. 2021 Apr 29;17(4):e1009073. doi: 10.1371/journal.ppat.1009073. PMID 33914852.
 204. Pinto G, Sampaio M, Dias O, Almeida C, Azeredo J, Oliveira H. Insights into the genome architecture and evolution of Shiga toxin encoding bacteriophages of *Escherichia coli*. *BMC Genomics*. 2021 May 19;22(1):366. doi: 10.1186/s12864-021-07685-0. PMID 34011288.
 205. Byrne L, Adams N, Jenkins C. Association between Shiga Toxin-Producing *Escherichia coli* O157:H7 *stx* Gene Subtype and Disease Severity, England, 2009-2019. *Emerg Infect Dis*. 2020 Oct;26(10):2394-2400. doi: 10.3201/eid2610.200319. PMID 32946720.
 206. Waters EV, Tucket LA, Ahmed JA, Wain J, Langridge GC. Impact of *Salmonella* genome rearrangement on gene expression. *Evol Lett*. 2022 Nov 19;6(6):426-437. doi: 10.1002/evl3.305. PMID 36579163.
 207. Fitzgerald SF, Lupolova N, Shaaban S, Dallman TJ, Greig DR, Allison L, tongue SC, Evans J, Henry MK, McNeilly TM, Bono JL, Gally DL. Genome structural variation in *Escherichia coli* O157:H7. *Microb Genom*. 2021 Nov;7(11):000682. doi: 10.1099/mgen.0.000682. PMID 34751643.
 208. West PT, Chanin RB, Bhatt AS. From genome structure to function: insights into structural variation in microbiology. *Curr Opin Microbiol*. 2022 Oct;69:102192. doi: 10.1016/j.mib.2022.102192. PMID 36030622.

209. Liu SL, Sanderson KE. Rearrangements in the genome of the bacterium *Salmonella typhi*. *Proc Natl Acad Sci USA*. 1995 Feb 14;92(4):1018-22. doi: 10.1073/pnas.92.4.1018. PMID 7862625.
210. Sanderson KE, Liu SL. Chromosomal rearrangements in enteric bacteria. *Electrophoresis*. 1998 Apr;19(4):569-72. doi: 10.1002/elps.1150190417. PMID 9588803.
211. Beutin L, Kaulfuss S, Herold S, Oswald E, Schmidt H. Genetic analysis of enteropathogenic and enterohemorrhagic *Escherichia coli* serogroup O103 strains by molecular typing of virulence and housekeeping genes and pulsed-field gel electrophoresis. *J Clin Microbiol*. 2005 Apr;43(4):1552-63. doi: 10.1128/JCM.43.4.1552-1563.2005. PMID 15814965.
212. Repar J, Supek F, Klanjsek T, Warnecke T, Zahradka K, Zahradka D. Elevated Rate of Genome Rearrangements in Radiation-Resistant Bacteria. *Genetics*. 2017 Apr;205(4):1677-1689. doi: 10.1534/genetics.116.196154. PMID 28188144.
213. Ring N, Abrahams JS, Jain M, Olsen H, Preston A, Bagby S. Resolving the complex *Bordetella pertussis* genome using barcoded nanopore sequencing. *Microb Genom*. 2018 Nov;4(11):e000234. doi: 10.1099/mgen.0.000234. PMID 30461375.
214. Seferbekova Z, Zabelkin A, Takovleva Y, Afasizhev R, Dranenko NO, Alexeev N, Gelfand MS, Bochkareva OO. High Rates of Genome Rearrangements and Pathogenicity of *Shigella* spp. *Front Microbiol*. 2021 Apr 12;12:628622. doi: 10.3389/fmicb.2021.628622. PMID 33912145.
215. Gilchrist CLM, Chooi YH. clinker & clustermap.js: automatic generation of gene cluster comparison figures. *Bioinformatics*. 2021 Aug 25;37(16):2473-2475. doi: 10.1093/bioinformatics/btab007. PMID 33459763.
216. Marçais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, Zimin A. MUMmer4: A fast and versatile genome alignment system. *PLoS Comput Biol*. 2018 Jan 26;14(1):e1005944. doi: 10.1371/journal.pcbi.1005944. PMID 29373581.
217. De Paepe M, Leclerc M, Tinsley CR, Petit MA. Bacteriophages: an underestimated role in human and animal health? *Front Cell Infect Microbiol*. 2014 Mar 28;4:39. doi: 10.3389/fcimb.2014.00039. PMID 24734220.
218. Lesterlin C, Pages C, Dubarry N, Dasgupta S, Cornet F. Asymmetry of chromosome Replichores renders the DNA translocase activity of FtsK essential for cell division and cell shape maintenance in *Escherichia coli*. *PLoS Genet*. 2008 Dec;4(12):e1000288. doi: 10.1371/journal.pgen.1000288. PMID 19057667.
219. Kuzminov A. Recombinational repair of DNA damage in *Escherichia coli* and bacteriophage lambda. *Microbiol Mol Biol Rev*. 1999 Dec;63(4):751-813. doi: 10.1128/MMBR.63.4.751-813.1999. PMID 10585965.
220. Blankenship HM, Dietrich SE, Burgess E, Wholehan J, Soehnlen M, Manning S. Whole-genome sequencing of Shiga toxin-producing *Escherichia coli* for characterization and outbreak investigation. *Microorganisms*. 2023 May 16;11(5):1298. doi: 10.3390/microorganisms11051298. PMID 37317272.
221. Nouws S, Verhaegen B, Denayer S, Crombé F, Piérard D, Bogaerts B, Vanneste K, Marchal K, Roosens NHC, De Keersmaecker SCJ. Transforming Shiga toxin-producing *Escherichia coli* surveillance through whole genome sequencing in food safety practices. *Front Microbiol*. 2023 Jul 13;14:1204630. doi: 10.3389/fmicb.2023.1204630. PMID 37520372.
222. Boyd EF. Bacteriophage-encoded bacterial virulence factors and phage-pathogenicity island interactions. *Adv Virus Res*. 2012;82:91-118. doi: 10.1016/B978-0-12-394621-8.00014-5. PMID 22420852.
223. Baker KS, Dallman TJ, Ashton PM, Day M, Hughes G, Crook PD, Gilbert VL, Zittermann S, Allen VG, Howden BP, Tomita T, Valcanis M, Harris SR, Connor TR, Sintchenko V, Howard P, Borwn JD, Petty NK, Gouali M, Thanh DP, Keddy KH, Smith AM, Talukder KA, Faruque SM, Parkhill J, Baker S, Weill FX, Jenkins C, Thomson NR. Intercontinental dissemination of azithromycin-resistant shigellosis through sexual transmission: a cross-sectional study.

- Lancet Infect Dis.* 2015 Aug;15(8):913-21. doi: 10.1016/S1473-3099(15)00002-X. PMID 25936611.
224. Ingle DJ, Nair S, Hartman H, Ashton PM, Dyson ZA, Day M, Freedman J, Chattaway MA, Holt KE, Dallman TJ. Informal genomic surveillance of regional distribution of *Salmonella* Typhi genotypes and antimicrobial resistance via returning travellers. *PLoS Negl Trop Dis.* 2019 Sep 12;13(9):e0007620. doi: 10.1371/journal.pntd.0007620. PMID 31513580.
 225. Bardsley M, Jenkins C, Mitchell HD, Mikhail AFW, Baker KS, Foster K, Hughes G, Dallman TJ. Persistent transmission of shigellosis in England is associated with a recently emerged multidrug-resistant strain of *Shigella sonnei*. *J Clin Microbiol.* 2020 Mar 25;58(4):e01692-19. doi: 10.1128/JCM.01692-19. PMID 31969425.
 226. Boxall MD, Day MR, Greig DR, Jenkins C. Antimicrobial resistance profiles of diarrhoeagenic *Escherichia coli* isolated from travellers returning to the UK, 2015-2017. *J Med Microbiol.* 2020 Jul;69(7):932-943. doi: 10.1099/jmm.0.001214. PMID 32530393.
 227. Nair S, Chattaway M, Langridge GC, Gentle A, Day M, Ainsworth EV, Mohamed I, Smith R, Jenkins C, Dallman TJ, Godbole G. ESBL-producing strains isolated from imported cases of enteric fever in England and Wales reveal multiple chromosomal integrations of *bla*_{CTX-M-15} in XDR *Salmonella* Typhi. *J Antimicrob Chemother.* 2021 May 12;76(6):1459-1466. doi: 10.1093/jac/dkab049. PMID 33704480.
 228. Day M, Doumith M, Jenkins C, Dallman TJ, Hopkins KL, Elson R, Godbole G, Woodford N. Antimicrobial resistance in Shiga toxin-producing *Escherichia coli* serogroups O157 and O26 isolated from human cases of diarrhoeal disease in England, 2015. *J Antimicrob Chemother.* 2017 Jan;72(1):145-152. doi: 10.1093/jac/dkw371. PMID 27678285.
 229. Gentle A, Day MR, Hopkins KL, Godbole G, Jenkins C. Antimicrobial resistance in Shiga toxin-producing *Escherichia coli* other than serotype O157:H7 in England, 2014-2016. *J Med Microbiol.* 2020 Mar;69(3):379-386. doi: 10.1099/jmm.0.001146. PMID 32101158.
 230. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.* 2021 Jul 2;49(W1):W293-W296. doi: 10.1093/nar/gkab301. PMID 33885785.
 231. Willshaw GA, Cheasty T, Smith HR, O'Brien SJ, Adak GK. Verocytotoxin-producing *Escherichia coli* (VTEC) O157 and other VTEC from human infections in England and Wales: 1995-1998. *J Med Microbiol.* 2001 Feb;50(2):135-142. doi: 10.1099/0022-1317-50-2-135. PMID 11211220.
 232. Threlfall EJ, Ward LR, Frost JA, Willshaw GA. The emergence and spread of antibiotic resistance in food-borne bacteria. *Int J Food Microbiol.* 2000 Dec 5;62(1-2):1-5. doi: 10.1016/s0168-1605(00)00351-2. PMID 11139009.
 233. Baker KS, Dallman TJ, Field N, Childs T, Mitchell H, Day M, Weill FX, Lefèvre S, Tourdjman M, Hughes G, Jenkins C, Thomson N. Genomic epidemiology of *Shigella* in the United Kingdom shows transmission of pathogen sublineages and determinants of antimicrobial resistance. *Sci Rep.* 2018 May 9;8(1):7389. doi: 10.1038/s41598-018-25764-3. PMID 29743642.
 234. Mitchell HD, Whitlock G, Zdravkov J, Olsson J, Silalang O, Bardsley M, Blomquist PB, Jenkins C, Thomson NR, Field N, Hughes G. Prevalence and risk factors of bacterial enteric pathogens in men who have sex with men: A cross-sectional study at the UK's largest sexual health service. *J Infect.* 2023 Jan;86(1):33-40. doi: 10.1016/j.jinf.2022.10.033. PMID 36343788.
 235. Thorley K, Charles H, Greig DR, Prochazka M, Mason LCE, Baker KS, Godbole G, Sinka K, Jenkins C. Emergence of extensively drug-resistant and multidrug-resistant *Shigella flexneri* serotype 2a associated with sexual transmission among gay, bisexual, and other men who have sex with men, in England: a descriptive epidemiological study. *Lancet Infect Dis.* 2023 Jun;23(6):732-739. doi: 10.1016/S1473-3099(22)00807-6. PMID 36731481.

236. Ni Y, Liu X, Simeneh ZM, Yang M, Li R. Benchmarking of Nanopore R10.4 and R9.4.1 flow cells in single-cell whole-genome amplification and whole-genome shotgun sequencing. *Comput Struct Biotechnol J*. 2023 Mar 24;21:2352-2364. doi: 10.1016/j.csbj.2023.03.038. PMID 37025654.
237. Wick RR, Holt KE. Benchmarking of long-read assemblers for prokaryote whole genome sequencing. *F1000Res*. 2019 Dec 23;8:2138. doi: 10.12688/f1000research.21782.4. PMID 31984131.
238. Yuen ZWS, Srivastava A, Daniel R, McNevin D, Jack C, Eyraas E. Systematic benchmarking of tools for CpG methylation detection from nanopore sequencing. *Nat Commun*. 2021 Jun 8;12(1):3438. doi: 10.1038/s41467-021-23778-6. PMID 34103501.

Chapter 8: Appendix

8.1 Data availability.

Illumina and Nanopore-based FASTQ files can be found on the National Center for Biotechnology information (NCBI) Sequence Read Archive (SRA) associated with BioProject ID PRJNA315192.

Illumina FASTQ files, Nanopore FASTQ files and Nanopore-based assemblies are linked and can be found under each BioSample ID within the above BioProject ID.

8.2 Software availability.

All publicly available software is referenced accordingly and can be obtained from said publication.

Software designed for this thesis specifically is available upon request.

Pipelines/workflows developed are in development and reserved for a separate publication (at time of writing).

8.3 Published materials from PhD analyses.

Manuscript 1 – Comparison of single-nucleotide variants identified by Illumina and Oxford Nanopore technologies in the context of a potential outbreak of Shiga toxin–producing *Escherichia coli*.

NCBI PubMed link – <https://pubmed.ncbi.nlm.nih.gov/31433830/>

Academic journal link – <https://doi.org/10.1093/gigascience/giz104>

Manuscript 2 – Comparison of Shiga toxin-encoding bacteriophages in highly pathogenic strains of Shiga toxin-producing *Escherichia coli* O157:H7 in the UK.

NCBI PubMed link – <https://pubmed.ncbi.nlm.nih.gov/32100710/>

Academic journal link – <https://doi.org/10.1099/mgen.0.000334>

Manuscript 3 – Analysis Shiga Toxin-Encoding Bacteriophage in a Rare Strain of
Shiga Toxin-Producing *Escherichia coli* O157:H7 *stx2a/stx2c*.

NCBI PubMed link – <https://pubmed.ncbi.nlm.nih.gov/33193199/>

Academic journal link – <https://doi.org/10.3389/fmicb.2020.577658>

Manuscript 4 – A Shiga Toxin-Encoding prophage recombination event confounds
the phylogenetic relationship between two isolates of *Escherichia coli*
O157:H7 from the same patient.

NCBI PubMed link – <https://pubmed.ncbi.nlm.nih.gov/33193248/>

Academic journal link – <https://doi.org/10.3389/fmicb.2020.588769>

Manuscript 5 – Analysis of a small outbreak of Shiga toxin-producing *Escherichia coli*
O157:H7 using long-read sequencing.

NCBI PubMed link – <https://pubmed.ncbi.nlm.nih.gov/33683192/>

Academic journal link – <https://doi.org/10.1099/mgen.0.000545>

Manuscript 6 – Re-analysis of an outbreak of Shiga toxin-producing *Escherichia coli*
O157:H7 associated with raw drinking milk using Nanopore sequencing.

NCBI PubMed link – <https://pubmed.ncbi.nlm.nih.gov/38461188/>

Academic journal link – <https://www.nature.com/articles/s41598-024-54662-0>

Manuscript 7 – Surveillance of antimicrobial resistant Shiga toxin-*producing E. coli*

O157:H7 in England, 2016–2020.

NCBI PubMed link – <https://pubmed.ncbi.nlm.nih.gov/37545157/>

Academic journal link – <https://academic.oup.com/jac/article/78/9/2263/7237980>



RESEARCH

Comparison of single-nucleotide variants identified by Illumina and Oxford Nanopore technologies in the context of a potential outbreak of Shiga toxin-producing *Escherichia coli*

David R. Greig , Claire Jenkins , Saheer Gharbia and Timothy J Dallman

National Infection Service, Public Health England, 61 Colindale Avenue, London NW9 5EQ, UK

*Correspondence address: Timothy J Dallman, National Infection Service, Public Health England, London NW9 5EQ, UK. E-mail: Tim.Dallman@phe.gov.uk <http://orcid.org/0000-0001-7105-2543>

Abstract

Background: We aimed to compare Illumina and Oxford Nanopore Technology sequencing data from the 2 isolates of Shiga toxin-producing *Escherichia coli* (STEC) O157:H7 to determine whether concordant single-nucleotide variants were identified and whether inference of relatedness was consistent with the 2 technologies. **Results:** For the Illumina workflow, the time from DNA extraction to availability of results was ~40 hours, whereas with the ONT workflow serotyping and Shiga toxin subtyping variant identification were available within 7 hours. After optimization of the ONT variant filtering, on average 95% of the discrepant positions between the technologies were accounted for by methylated positions found in the described 5-methylcytosine motif sequences, CC(A/T)GG. Of the few discrepant variants (6 and 7 difference for the 2 isolates) identified by the 2 technologies, it is likely that both methodologies contain false calls. **Conclusions:** Despite these discrepancies, Illumina and Oxford Nanopore Technology sequences from the same case were placed on the same phylogenetic location against a dense reference database of STEC O157:H7 genomes sequenced using the Illumina workflow. Robust single-nucleotide polymorphism typing using MinION-based variant calling is possible, and we provide evidence that the 2 technologies can be used interchangeably to type STEC O157:H7 in a public health setting.

Keywords: Oxford Nanopore; Illumina; variant calling; STEC; outbreak

Background

Shiga toxin-producing *Escherichia coli* (STEC) O157:H7 is a zoonotic, foodborne pathogen defined by the presence of phage-encoded Shiga toxin genes (*stx*) [1]. Disease symptoms range from mild through to severe bloody diarrhoea, often accompanied by fever, abdominal cramps, and vomiting [2]. The infection can progress to haemolytic uraemic syndrome (HUS), characterized by kidney failure and/or cardiac and neurological complications [3, 4]. Transmission from an animal reservoir, mainly ruminants, occurs by direct contact with animals or their environment, or by the consumption of contaminated food

products, with reported vehicles including beef and lamb meat, dairy products, raw vegetables, and salad [2, 4].

STEC O157:H7 belongs to multi-locus sequence type (MLST) clonal complex (CC) 11, with all but a small number of variants belonging to sequence type (ST) 11. CC11 comprises 3 main lineages (I, II, and I/II) and 7 sublineages (Ia, Ib, Ic, IIa, IIb, IIc, and I/II) [5]. There are 2 types of Shiga toxin, Stx1 and Stx2. Stx1 has 4 subtypes (1a–1d) and Stx2 has 7 subtypes (2a–2g). Subtypes 1a, 2a, 2c, and rarely 2d are found in STEC O157:H7. Strains harbouring *stx2a* are significantly associated with cases that develop HUS [2, 6]. As well as harbouring *stx* encoding prophage, STEC

Received: 6 March 2019; Revised: 12 June 2019; Accepted: 7 August 2019

© The Author(s) 2019. Published by Oxford University Press. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

O157:H7 has an additional prophage repertoire accounting for $\geq 20\%$ of the chromosome.

The implementation of whole-genome sequencing data for typing STEC has improved the detection and management of outbreaks of foodborne disease [6]. Single-nucleotide polymorphism (SNP) typing offers an unprecedented level of strain discrimination and can be used to quantify the genetic relatedness between groups of genomes. In general, for clonal bacteria, the fewer polymorphisms identified between pairs of strains, the less time since divergence from a common ancestor and therefore the increased likelihood that they are from the same source population. Therefore, it is paramount that variant detection for typing be accurate, highly specific, and concentrated on positions of neutral evolution to ensure the correct interpretation of the sequence data within the epidemiological context of an outbreak. It has been previously shown that different bioinformatics analysis approaches for variant identification exhibit detection variability [7, 8]. It is therefore important that, within a particular analysis, workflow parameters to filter identified variants to achieve optimum sensitivity and specificity are appropriately optimized.

Short-read sequencing platforms, such as those provided by Illumina, have been adopted by public health agencies for infectious disease surveillance worldwide [9] and have proved to be a robust and accurate method for quantifying relatedness between bacterial genomes. High-throughput Illumina sequencing, although cost-effective, often requires batch processing of hundreds of microbial isolates to achieve cost savings and therefore this approach offers less flexibility for urgent, small-scale sequencing often required during public health emergencies [10]. In contrast, Oxford Nanopore Technologies (ONT) offers a range of rapid real-time sequencing platforms from the portable MinION to the higher throughput GridION and PromethION models, although at this time lower read accuracy compared with Illumina data suggests that accurate variant calling may be problematic.

In September 2017, Public Health England (PHE) was notified of 2 cases of HUS in 2 children admitted to the same hospital on the same night. STEC O157:H7 was isolated from the faecal specimens of both cases. To rapidly determine whether the cases were part of a related phylogenetic cluster and therefore likely to be epidemiologically linked to each other or to any other cases in the PHE database, we sequenced both isolates using the MinION platform and integrated the ONT sequencing data with a dense reference database of Illumina sequences. We aimed to compare Illumina and ONT sequencing data from the 2 isolates to assess the utility of the ONT method for urgent, small-scale sequencing and to determine whether the same single-nucleotide variants were identified and whether inference of relatedness was consistent with the 2 technologies.

Data Description

Paired-end FASTQ files were generated from the Illumina HiSeq 2500 for both samples (cases). Raw long-read data (FAST5) were generated from the MinION and base-called using Albacore (FASTQ) in real time. Both technologies' derived FASTQ reads were trimmed and filtered (Trimmomatic, Porechop, Filtlong) before being aligned (BWA, Minimap2) to a reference genome (NC_002695.1). Variant positions were called using GATK before being imported into SnapperDB. Full processing details can be found in the Methods section.

Results

Comparison of typing results generated by Illumina and ONT workflows

To consider the potential benefits of real-time sequencing to enhance opportunities for early outbreak detection, the timelines from DNA extraction to result generation for Illumina and ONT workflows were evaluated (Fig. 1) and the relationship between yield, time, and genome coverage plotted (Fig. 2). For the ONT workflow, the time from DNA extraction to completion of the sequencing run was 28 hours. A total yield of 0.45 Gb for the isolate from Case A and 0.59 Gb for the isolate from Case B was achieved, which corresponds to an equivalent coverage of the Sakai O157 STEC reference genome (5.4 Mb) of 81.29 $\times$ and 108.30 $\times$ for isolate A and B, respectively. The mean PHRED quality score for all reads in Case A was 9.87 and Case B was 9.47, which is ~ 1 error every 10 bases. Base-calling and analysis was performed in real time, and serotyping, Shiga toxin subtyping, and variant identification were available within 6 hours and 20 minutes of the 24-hour sequencing run. With respect to the Illumina sequencing workflow, the time from DNA extraction to availability of results, assuming there were no breaks in the process, was just under 40 hours (Fig. 1).

The species identification, serotype, MLST profile, and Shiga toxin subtype results generated by both Illumina and ONT workflows were concordant with both isolates identified as *E. coli* O157:H7 ST11, *stx2a* and *stx2c*. During the ONT sequencing run, the bacterial species was unambiguously identified in < 1 minute for both cases (Fig. 1). Additionally, using Krocus, a confirmed MLST was generated for Case A at 1:54 hours and Case B at 10:39 hours into the sequencing run. This was the point at which the last read required to generate a consensus on the MLST was base-called. By 93 minutes for Case A and 41 minutes for Case B, it was possible to determine the *E. coli* O157:H7 serotype, and *stx2a* and *stx2c* were detected at 58 and 24 minutes into the sequencing run for Case A and Case B, respectively.

Optimization of ONT variant calling

To compare Illumina and ONT sequences within a standardized framework it was necessary to optimize the parameters for variant filtering within GATK2 to compensate for the lower read accuracy observed in the ONT data. Using Case B for the optimization, base calls in the ONT data were classified as true-positive (variant base detected by both methods), false-positive (variant base in ONT, reference base in Illumina), true-negative (reference base in Illumina and ONT), or false-negative results (variant base in Illumina, reference base in ONT). To disregard areas of the genome that the ONT reads could map to (and therefore identify variants) but were ambiguously mapped with Illumina reads, prefiltering was performed by masking regions annotated as phage in the reference genome and those that could not be accurately self-mapped with simulated reference Illumina FASTQ reads. Fig. 3 plots the precision (the proportion of true-positive results with respect to all positive calls) against the recall/sensitivity (the proportion of true-positive results identified with respect to all true-positive results) for an array of consensus ratio cut-offs for each of the masking strategies. Similar areas under the curve (AUCs) were achieved for the different masking strategies, with slightly higher precision at lower recall achieved with "self-masking" (AUC, 0.71) and slightly higher recall at lower precision with explicit masking

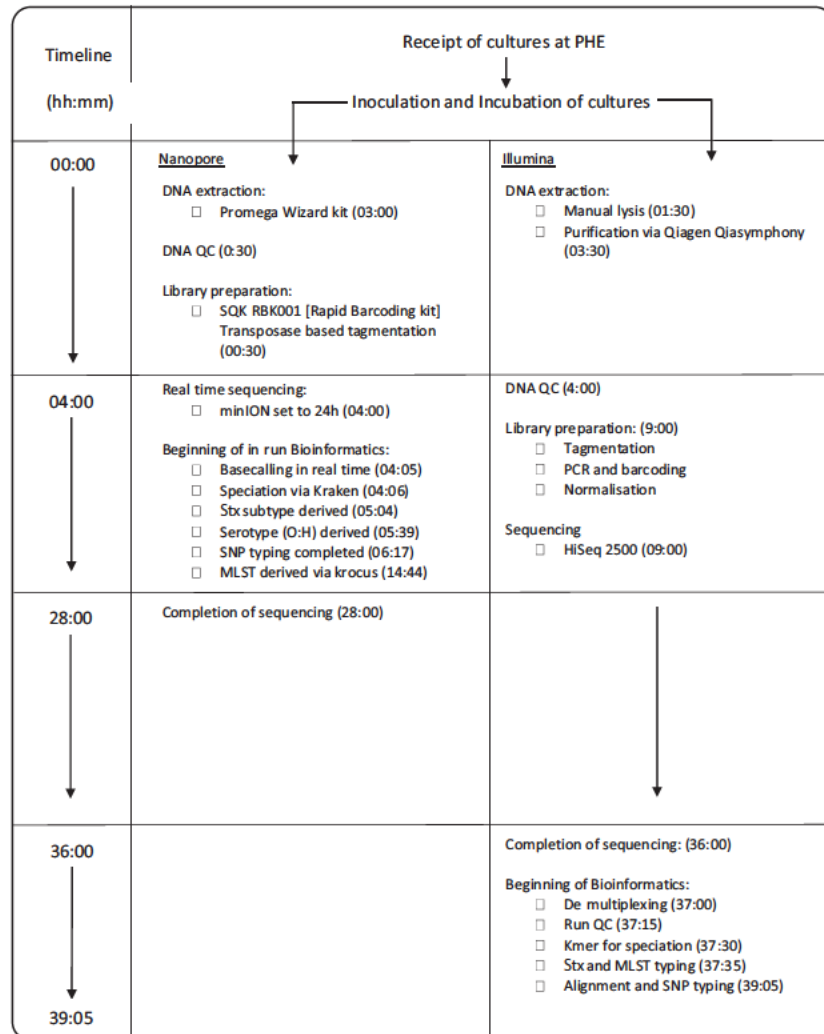


Figure 1: Figure showing comparative timeline from beginning DNA extraction to results generation for Oxford Nanopore and Illumina technologies. Times show the completion of the labelled event relative to the start of the assay (hh:mm). QC: quality control; Stx: Shiga toxin.

of the Sakai prophage (AUC, 0.75). The absence of a masking strategy markedly affects the precision of variant calling with ONT data, in comparison with Illumina as gold standard (AUC, 0.30). To identify the optimum consensus cut-off for filtering ONT variants processed through GATK the F1 score was calculated at each consensus cut-off. A consensus cut-off of 0.8 maximized the precision and recall (Fig. 4) irrespective of the filtering methods.

Investigation of the discrepant variants identified between the Illumina and ONT data

After optimized quality and prophage filtering there were 266 and 101 base positions for Cases A and B, respectively, that were discordant between the ONT and Illumina sequencing data. The majority of discrepancies were where the ONT data identified a variant not identified in the Illumina data (261 of 266 [98.1%] and 95 of 101 [94.1%] discrepant base positions for Cases A and

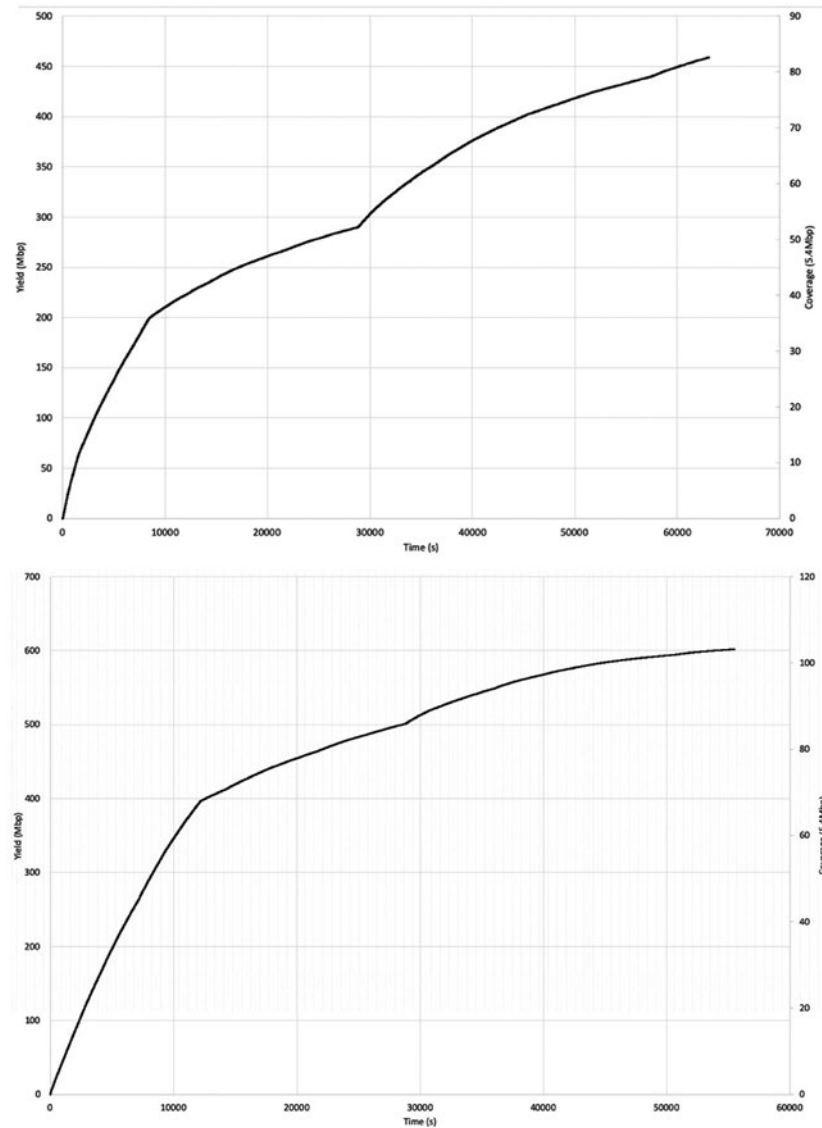


Figure 2: Time/yield/coverage graphs showing production of reads in real time and the associated cumulative mapping coverage. Case A is the graph on the top and Case B is below.

B, respectively). In contrast the Illumina data identified 5 (1.9%) discrepant base positions as variants for Case A and 6 (5.9%) for case B (Table 1) not identified by the ONT data.

For both cases the most common discrepant variants were adenines classified as guanines in the ONT data with respect to the Illumina data (and reference), accounting for 68.0% (181/266)

for Case A and 72.3% (73/101) for Case B. The second most common discrepancy was thymine being classified as cytosine in the ONT data accounting for 29.7% (79/266) in Case A and 20.8% (21/101) in Case B (Table 1). Of the transitions described above, 97.7% (Case A) and 93.1% (Case B) occurred when the variant was between 2 homopolymeric regions of multiple cy-

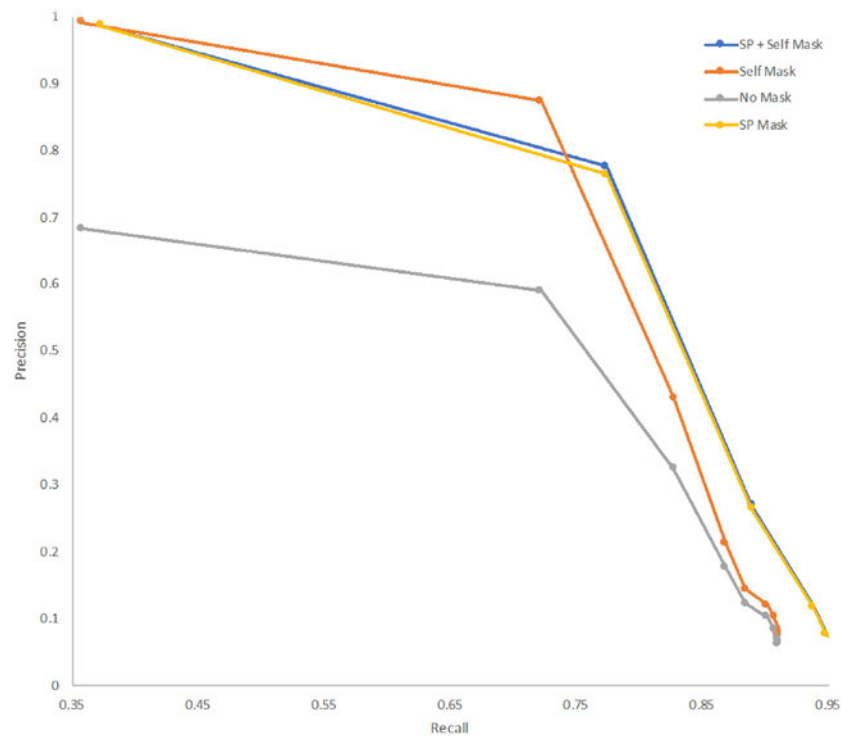


Figure 3: Precision vs recall of variant calling for an array of consensus ratio cut-offs and premasking strategies including masking positions annotated as “Sakai phage” (“SP”) and positions that are ambiguously self-mapped (“Self”) with simulated Illumina FASTQs from the reference genome. Performed on case B.

Table 1: Breakdown of the total number of variants of each technology against the reference genome, followed by the numbers of masked variants within prophage or methylated positions

Variants and reason for omission	Case A		Case B	
Total No. of variants against the reference genome after quality filtering	2,076		1,424	
Total No. of variants with masked due to location in phage	708		531	
Total No. of discrepant variants called between Case A and B alone	266		101	
Variants and reason for omission	Illumina VCF	ONT VCF	Illumina VCF	ONT VCF
No. of discrepant variants in each VCF	5	261	6	95
No. of discrepant variants with methylated positions masked	0	260	0	94
Final discrepant variants	5	1	6	1

tosines and guanines (Fig. 5). These homopolymeric regions were similar to described DNA cytosine methylase (Dcm) binding sequences [11]. Nanopolish was subsequently used to identify likely Dcm, 5'-cytosine-phosphate-guanine-3' (CpG), and DNA adenine methyltransferase (Dam) methylation sites in the ONT sequencing data and confirmed that 260 of 266 (97.7%) and 94 of 101 (93.1%) discrepant variants in the ONT data were classed as methylated for Cases A and B, respectively.

All of these were determined to be Dcm methylation for both cases.

Once the methylated positions were masked from the analysis, there were a total of 6 (5 discrepant variants in Illumina and 1 ONT) and 7 (6 discrepant variants in Illumina and 1 ONT) discrepant SNPs between the ONT and Illumina data, for Cases A and B, respectively (Tables 2 and 3). Four discrepant Illumina variants are shared by both Case A and Case B. One shared vari-

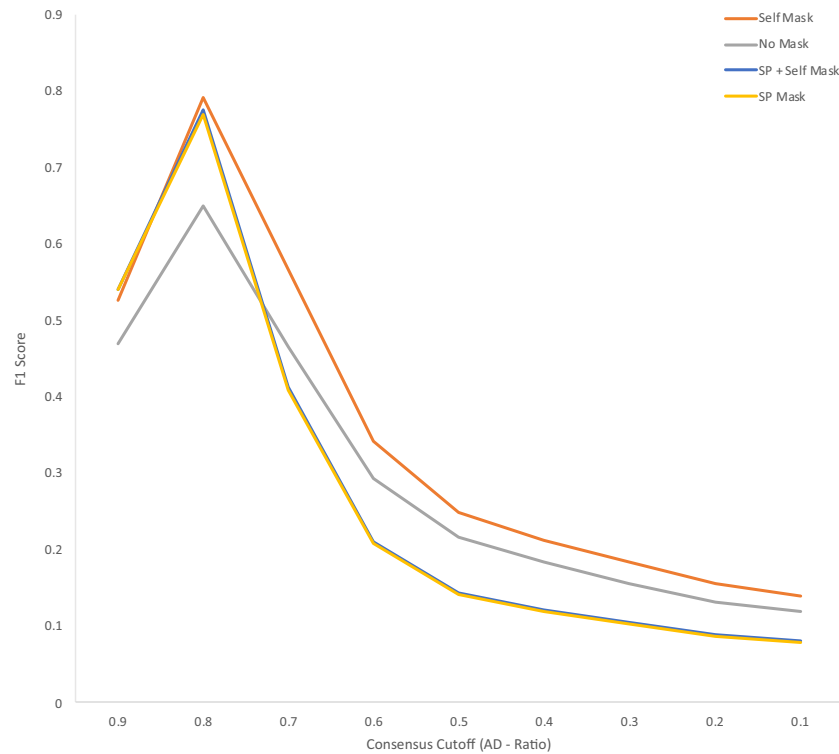


Figure 4: F1 Score for an array of consensus ratio cut-offs and premasking strategies including masking positions annotated as "Sakai phage" ("SP") and positions that are ambiguously self-mapped ("Self") with simulated Illumina FASTQs from the reference genome.

Table 2: Final discrepant SNPs between the Illumina data and ONT data for Case A

SNP	Position	Base in reference	Base in Illumina	Depth in Illumina	Base in ONT	Depth in ONT	Variant	Locus tag	Annotation
1	270,595	C	A	46	C	141	A	ECs0237	rhsC
2	379,516	A	G	114	A	100	G	NON CODING	N
3	1,681,338	C	G	59	C	61	G	ECs1685	Alanine racemase 2
4	1,681,339	G	C	57	G	61	C	ECs1685	Alanine racemase 2
5	2,636,513	T	C	91	T	69	C	ECs2674	Hypothetical protein
6	4,709,195	A	A	86	G	82	G	ECs4673	Membrane-bound ATP synthase epsilon-subunit AtpC

Also shown is the base as it is in the reference, the Illumina called base and read depth at that position, and the same for the ONT data. Finally, also included is the locus tag relative to the reference genome and the gene annotation.

ant was found in a non-coding region; another shared variant was found in *rhsC* encoding an RHS (rearrangement hotspot) protein defined by the presence of extended repeat regions. Two further shared variants were found in *dadX*, an alanine racemase gene. *dadX* is a paralogue of *alr*, also annotated as an alanine racemase in the Sakai reference genome with significant nucleotide similarity (>75% nucleotide identity). Both intra- and inter-gene repeats are known to be regions of potential false-positive calls with Illumina data due to mis-mapping. Of the 7

variants in the Illumina data found in either or both Case A and B, 5 were found to be homoplastic in the O157 population of 4,475 Illumina sequences, arising independently, multiple times.

Phylogenetic analysis

Using the optimized variant calling parameters both strains clustered phylogenetically in lineage 1c within a dense reference database of STEC O157:H7 genomes ($n = 4,475$). However, the

Table 3: Final discrepant SNPs between the Illumina data and ONT data for Case B

SNP	Position	Base in reference	Base in Illumina	Depth in Illumina	Base in ONT	Depth in ONT	Variant	Locus tag	Annotation
1	270,595	C	A	19	C	207	A	ECs0237	rhsC
2	379,516	A	G	52	A	124	G	NON CODING	N
3	1,681,338	C	G	44	C	86	G	ECs1685	Alanine racemase 2
4	1,681,339	G	C	41	G	86	C	ECs1685	Alanine racemase 2
5	2,033,176	T	G	34	T	85	G	ECs2049	Hypothetical protein
6	2,731,621	A	C	52	A	73	C	NON CODING	N
7	4,901,209	A	A	49	G	102	G	ECs4834	Superoxide dismutase SodA

Also shown is the base as it is in the reference, the Illumina called base and read depth at that position, and the same for the ONT data. Finally, also included is the locus tag relative to the reference genome and the gene annotation.

	Position					Case	
	-2	-1	Variant	+1	+2	A	B
Reference	C	C	A	G	G	A > G	69.62%
			↓			Transition	(n=181)
Alignment	C	C	G	=	G		(n=73)
Reference	C	C	T	G	G	T > C	30.38%
			↓			Transition	(n=79)
Alignment	C	C	C	G	G		(n=21)
						Total	100%
							(n=260)
							100%
							(n=94)

Figure 5: The 2 most common discrepancies in the ONT optimized GATK VCFs and a breakdown of the relative proportions of these transitions compared to the total number of discrepant SNPs for both cases.

genomes were located in distinct subclades (Fig. 6). It was, therefore, unlikely that the isolates originated from the same source, and it was concluded that Cases A and B were not epidemiologically linked. Following phylogenetic analysis of the Illumina SNP typing data (Fig. 6), Case A was designated a sporadic case. However, Case B clustered with a concurrent outbreak, already under investigation, comprising 3 additional cases. The Illumina sequence linked to Case B was zero SNPs different from the other 3 cases in the cluster, whereas the ONT sequence was 7 SNPs different, when excluding the methylated positions (Table 3). Based on the ONT sequencing data alone, this discrepancy would have led to uncertainty as to whether Case B was linked to the outbreak.

Assembly profile

The ONT-only assembly resolved to 5 contigs (5.73 Mb) for Case A and 4 contigs (5.60 Mb) for Case B (Supplementary Table S1). In Case A, the 5 contigs were determined to be a single chromosomal contig, a single plasmid contig (pO157), and the 3 prophage duplications. In Case B, the 4 contigs were determined to be a single chromosomal contig with 2 plasmids (1 being the pO157). For Case A the assembly resolved to 25 contigs (5.51 Mb) with a hybrid assembly and 668 contigs (5.45 Mb) with an Illumina-only assembly. Case B resolved to 34 contigs (5.49 Mb) with a hybrid assembly and 575 contigs (5.42 mb) with an Illumina-only assembly.

Alignment of the assemblies (Supplementary Figs S1 and S2) revealed several locations within the ONT-only assembly that were absent in the hybrid and Illumina-only assemblies. In Case A, there were 8 regions only present within the ONT-only chromosome assembly, of which 7 are related to prophage regions (Supplementary Fig. S1). In case B, there were 10 chromosomal

regions in the ONT-only assembly that did not align to the other assemblies. All 10 regions were associated with prophage regions (Supplementary Fig. S2).

Discussion

In this study, the 2 isolates sequenced using ONT were unambiguously identified as STEC O157:H7 ST 11 stx2a/stx2c in <15 hours and it was possible to distinguish the genetic relatedness between the isolates within 377 minutes (i.e., 22 hours before the ONT sequencing run was scheduled to finish and just under 3 hours before the Illumina sequencing began). The whole-genome sequencing turnaround time, from DNA extraction and library preparation to sequencing and analysis via the Illumina workflow at PHE, is 3–6 days. Although this turnaround time is rapid for a service using batch processing on the HiSeq platforms, the sequencing approach using the MinION, whereby individual samples or small barcoded batches are loaded and results generated and analysed in real time, has the potential to be faster and more flexible. It should be noted that speed to result for ONT sequencing will be related to the amount of isolates run on a single flow cell because DNA molecules from different samples will compete to traverse a finite number of pores. This approach is therefore ideal for urgent, small-scale sequencing, often required during public health emergencies. In this scenario, analysis of the ONT data provided evidence that the 2 cases were not epidemiologically linked and, although efforts were made to determine the potential source of the infection for both cases through the National Enhanced STEC Surveillance System [2], an outbreak investigation was not initiated.

A current limitation of MinION sequencing is its lower read accuracy when compared to short-read technologies [12–16]. This accuracy has improved as the technology has matured but still falls short of the 99% accuracy offered by short-read platforms [15]. There are a number of factors that contribute to the current read accuracy in the nanopore data including structural similarity of nucleotides, simultaneous influence of multiple nucleotides on the signal, the non-uniform speed at which nucleotides pass through the pore, and the fact that the signal does not change within homopolymers [15].

Although analysis of the Illumina and ONT sequencing data placed the sequences on the same branch on the phylogeny, there were SNP discrepancies between the sequences generated by the 2 different workflows, even after optimization of the parameters. The vast majority of the discrepant SNPs (261/266 [98.1%] and 95/101 [94.1%] for Cases A and B, respectively) were attributed to variants identified in the ONT data and not the Illumina data. The majority of discrepancies (97.7% in Case A and 93.1% in Case B) were found in sequences that are the same as

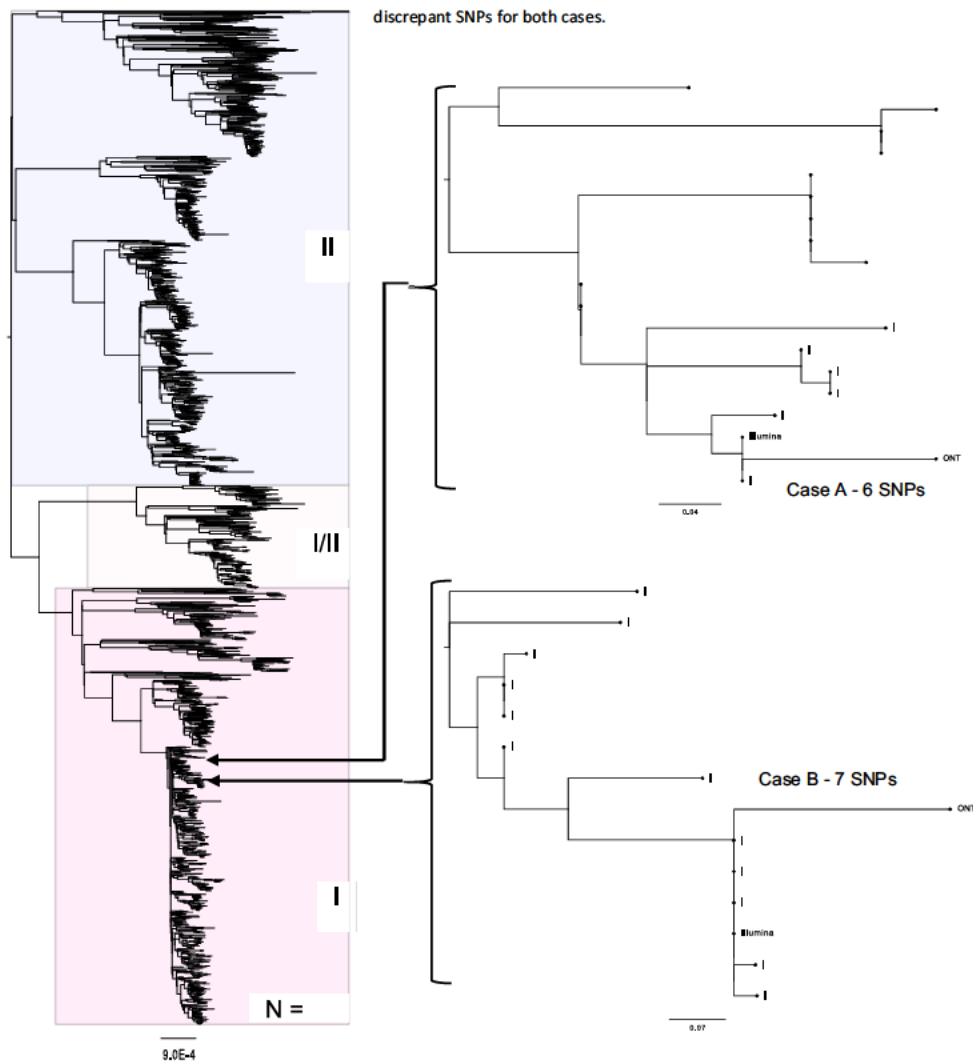


Figure 6: Maximum likelihood tree, of a "soft core" alignment of 4,475 genomes showing the tree lineages (I, I/II, and II) of STEC (Clonal Complex 11). Also showing where Oxford Nanopore and Illumina sequencing data are placed within the tree for each of the 2 cases. All methylated positions and prophage regions have been masked. Values represent the SNP differences between the Illumina and ONT data for both cases.

the known 5-methylcytosine motif sequences, CC(A/T)GG [11, 17] in the ONT data. Following a search of the ONT discrepant SNPs for CpG, Dam, and Dcm methylation using Nanopolish, the majority (97.7% and 93.1% for case A and B, respectively) of the ONT discrepant SNPs were identified in Dcm methylated regions.

As Nanopolish is detecting these methylated positions with the use of the raw FAST5 data, it is suggested that these particular discrepancies appear during the base-calling process. Albacore handles most methylation well across the 3 methylation models searched for by Nanopolish; e.g., only 94 of 13,504 methylated positions were considered incorrect by base calling

for Case B. However, for mapping-based SNP typing, this level of error in base calling means that it is not possible to accurately determine the number of SNPs, thus potentially obscuring the true phylogenetic relationship between isolates of STEC O157:H7.

The optimization of variant filtering was performed using the Illumina data as a gold standard. However, it is possible that the alignment of the Illumina data might have generated false SNPs based on reads mapping to ambiguous regions of the genome, whereas the long reads obtained using the ONT workflow are able to resolve these ambiguous regions and call variants, or not, at these positions correctly. Because the Illumina data were used as the gold standard, in this scenario SNPs produced in the Illumina data would have been classified incorrectly as false-negative results in the ONT data. Discrepant variants identified in the Illumina data were attributed mainly to potentially false mapping of Illumina reads to homologous regions of the reference genome, variants that were misidentified at the same position independently in Cases A and B. Furthermore, comparison of assemblies generated by ONT reads, Illumina reads, and a hybrid approach highlights the extra genetic content accessible to ONT assemblies where variation can be quantified.

In this study an ONT sequencing workflow was used to rapidly rule out an epidemiological link between 2 children admitted to the same hospital on the same day with symptoms of HUS. The isolates of STEC O157:H7 from each child mapped to different clades within the same STEC O157:H7 lineage (Ic). We provide further evidence that SNP typing using MinION-based variant calling is possible when the coverage of the variation is high [15]. The error rate exhibited by ONT sequencing workflows continues to improve as a result of developments in the pore design, the library preparation methods, innovations in base-calling algorithms, and the introduction of post-sequencing correction tools, such as Nanopolish [15, 18]. Currently, both short- and long-read technologies are used for public health surveillance, and there is a need to integrate the outputs so that all the data can be analysed in the same way. Recently, Rang et al. [15] reiterated how the scientific community can make valuable contributions to improving ONT read accuracy by systematically comparing computational strategies as highlighted in this study and elsewhere [19]. Ongoing updates to the chemistry and software tools will facilitate the robust detection of SNPs, enabling ONT to compete with short-read platforms, ultimately enabling the 2 technologies to be used interchangeably in clinical and public health settings.

Methods

DNA extraction, library preparation, and Illumina sequencing

Genomic DNA (gDNA) was extracted from 2 strains of STEC O157 isolates from 2 HUS cases admitted to the same hospital on the same night. Using a Qiagen Qiasymphony (Qiagen, Hilden, Germany) in accordance with the manufacturer's instructions, gDNA was extracted and quantified using a Qubit and the BR dsDNA Assay Kit (ThermoFisher Scientific, Waltham, MA, USA) to manufacturer's instructions. The sequencing library was prepared by fragmenting and tagging the purified gDNA using the Nextera XT DNA Sample Preparation Kits (Illumina, Cambridge, UK) to manufacturer's instructions. The prepared library was loaded onto an Illumina HiSeq 2500 (Illumina, Cambridge, UK) at PHE and sequencing performed in rapid run mode, yielding paired-end 100-bp reads.

Processing and analysis of Illumina sequence data

FASTQ reads were processed using Trimmomatic v0.27 (Trimmomatic, [RRID:SCR.011848](#)) [20] to remove bases with a PHRED score of <30 from the leading and trailing ends, with reads <50 bp after quality trimming discarded. A k-mer approach [21] was used to confirm the species of the samples. ST assignment was performed using MOST v1.0 described by Tewolde et al. [22]. *In silico* serotyping was performed by using GeneFinder [23], which uses Bowtie v2.2.5 (Bowtie, [RRID:SCR.005476](#)) [24] and Samtools v0.1.18 (SAMTOOLS, [RRID:SCR.002105](#)) [25] to align FASTQ reads to a multifasta containing the target genes (including *wzx*, *wzy*, and *fliC*). Stx subtyping was performed as described by Ashton et al. [26]. Illumina FASTQ reads were mapped to the Sakai STEC O157 reference genome (NC_002695.1) using BWA MEM v0.7.13 (BWA, [RRID:SCR.010910](#)) [27]. Variant positions were identified by GATK v2.6.5 UnifiedGenotyper (GATK, [RRID:SCR.001876](#)) [28] that passed the following parameters: >90% consensus, minimum read depth of 10, Mapping Quality (MQ) ≥ 30. Any variants called at positions that were within the known prophages in Sakai were masked from further analyses. The remaining variants were imported into SnapperDB v0.2.5 [29].

DNA extraction, library preparation, and nanopore sequencing

To preserve DNA integrity for the nanopore sequencing, gDNA was extracted and purified using the Promega Wizard Genomic DNA Purification Kit (Promega, Madison, WI, USA) with minor alterations including doubled incubation times, no vigorous mixing steps (performed by inversion), and elution into 50 µL of double-processed nuclease-free water (Sigma-Aldrich, St. Louis, MO, USA). DNA was quantified using a Qubit and the HS (High Sensitivity) dsDNA Assay Kit (ThermoFisher Scientific) according to the manufacturer's instructions. Library preparation was performed using the Rapid Barcoding Kit—SQK-RBK001 (ONT, Oxford, UK) with each sample's gDNA being barcoded by transposase-based tagmentation and pooled as per manufacturer's instructions. The prepared library was loaded on a FLO-MIN106 R9.4 flow cell (ONT, Oxford, UK) and sequenced using the MinION for 24 hours.

Processing and analysis of nanopore sequence data

Raw FAST5 files were base-called and de-multiplexed in real time, as reads were being generated, using Albacore v2.1 (ONT) into FASTQ files. Run metrics were generated using Nanoplot v1.8.1 using default parameters [30]. Reads were processed through Porechop v0.2.1 using default parameters (R Wick, unpublished results) [31] to remove any barcodes and adaptors used in SQK-RBK001. For Case A 96,788 reads (10,214,353 bases) were adaptor trimmed and 386 (0.39%) chimeric reads split. For Case B 430,911 reads (34,888,999 bases) were adaptor trimmed and 513 (0.11%) chimeric reads split. Samples were speciated using Kraken v0.10.4 [32]. An MLST was assigned using Krocus with the following parameters: *-kmer* 15, *-min_block_size* 300, and *-margin* 500 [33]. Stx subtyping and serotyping was determined by aligning the base-called reads using minimap2 v2.2 [34] and Samtools v1.1 [25] to a multifasta containing the Stx and serotype encoding genes.

For reference-based variant calling FASTQ reads were mapped to the Sakai STEC O157 reference genome (NC_002695.1) using minimap2 v2.2 [34]. VCFs were produced using GATK v2.6.5 UnifiedGenotyper [28]. Any variants called at positions that were

within the known prophages in Sakai were masked from further analyses. To determine the optimum consensus cut-off for ONT variant detection the VCF was filtered with sequentially decreasing ad-ratio values at 0.1 intervals. Using the Illumina variant calls as the gold standard, F1 scores (the weighted average of precision and recall) were calculated to determine the optimal ad-ratio for processing ONT data through GATK.

Comparison of Illumina and Nanopore discrepant SNPs

Nanopolish [18] was also used to detect methylation across the ONT data to compare to the discrepant positions. This was performed using the call-methylation function, searching for 3 types of methylation including the DNA adenine methyltransferase (Dam), DNA cytosine methylase (Dcm), and 5'-cytosine-phosphate-guanine-3' (CpG) models. The discrepant SNPs between the Illumina and ONT for both Case A and Case B were manually visualized in Tablet v1.17.08.17 [35] in order to elucidate the reason for the discrepancy. Discordant SNPs found within a homopolymeric region were also quantified.

Generation of phylogenetic trees

Filtered VCF files for each of the Illumina and ONT sequencing data for each sample were incorporated into SnapperDB v0.2.5 [29] containing variant calls from 4,471 other STEC CC11 genomes generated through routine surveillance by PHE. SnapperDB v0.2.5 [29] was used to generate a whole-genome alignment of the 4,475 genomes (including both datasets for the selected strains for this study). Both methylated positions and prophage positions were masked from the alignment. The alignment was processed through Gubbins V2.0.0 [36] to account for recombination events. A maximum likelihood tree was then constructed using RAxML V8.1.17 [37].

Assembly of ONT data

Trimmed ONT FASTQ files were assembled using Canu v1.6 (Canu, [RRID:SCR_015880](#)) [38]. Polishing of the assemblies was performed using Nanopolish v0.10.2 [18] using both the trimmed ONT FASTQs and FAST5s for each respective sample, accounting for methylation using the `-methylation-aware` option set to `dcm`. Assemblies were reoriented to start at the *dnaA* gene (NC_000913) from *E. coli* K12, using the `fixstart` parameter in circulator v1.5.5 [39].

Hybrid assemblies

Trimmed ONT FASTQ files were assembled using Unicycler v0.4.2 [40] with the following parameters: `min.fasta.length = 1000`, `mode = normal` and `-1` and `-2` for the incorporation of each sample's equivalent Illumina FASTQ. Pilon v1.23 [41] was used to correct the assembly using the Illumina reads.

Assembly of Illumina data

Illumina reads were assembled using SPAdes v3.13.0 (SPAdes, [RRID:SCR_000131](#)) [42] with the careful parameter activated and with *k*-mer lengths of 21, 33, 55, 65, 77, 83, and 91.

Annotation

Prokka v1.13 [43] with the species set to *E. coli* was used to annotate the final assemblies.

Mauve snapshot.2015-02-25 (1) [44] using the "move contig" function was used to align each assembly to the ONT reference as they had the fewest contigs.

Availability of supporting data and materials

The FASTQ files for the paired-read Illumina sequence data can be found on the NCBI SRA; Case A accession: SRR7184397, Case B accession: SRR6052929. The ONT FASTQ files, Case A accession: SRR7477814, Case B accession: SRR7477813. All files can be found under BioProject—PRJNA315192. The assemblies in the supplementary data were submitted to NCBI GenBank and are found under the following accessions: Nanopore assemblies (BioProject - PRJNA548323): ONT Case A: VIFR000000000, ONT Case B: VIFQ000000000. Illumina assemblies (BioProject - PRJNA315192): Illumina Case A: VIFT000000000, Illumina Case B: VIFS000000000. Hybrid assemblies (BioProject - PRJNA548322): Hybrid Case A: VIFP000000000, Hybrid Case B: VIFO000000000. All additional supporting data are available in the GigaScience GigaDB repository [45].

Additional files

Supplementary figure 1—Mauve alignment showing regions of similarity between the ONT-only, hybrid, and Illumina-only assemblies (order descending) for Case A. Also showing the chromosomal regions in the ONT-only assembly that did not match the other assemblies (red arrows).

Supplementary figure 2—Mauve alignment showing regions of similarity between the ONT-only, hybrid, and Illumina-only assemblies (order descending) for Case B. Also showing the chromosomal regions in the ONT-only assembly that did not match the other assemblies (red arrows).

Supplementary Table 1—Number of contigs generated and size of assembly for each assembly method for both cases.

Abbreviations

AUC: area under curve; bp: base pairs; BWA: Burrows-Wheeler aligner; CC: clonal complex; Dam: DNA adenine methyltransferase; Dcm: DNA cytosine methylase; GATK: Genome Analysis Toolkit; Gb: gigabase pairs; gDNA: genomic DNA; HUS: haemolytic uraemic syndrome; Mb: megabase pairs; MLST: multi-locus sequence type; NCBI: National Center for Biotechnology Information; ONT: Oxford Nanopore Technologies; PHE: Public Health England; RAxML: Randomized Axelerated Maximum Likelihood; SNP: single-nucleotide polymorphism; SPAdes: St. Petersburg genome assembler; SRA: Sequence Read Archive; ST: sequence type; STEC: Shiga toxin-producing *Escherichia coli*; VCF: variant call format.

Competing interests

This project was part funded by Oxford Nanopore Technologies.

Authors' contributions

C.J. and T.J.D. conceptualized the project. C.J. and D.R.G. performed DNA extractions. D.R.G. performed library preparation and Nanopore sequencing. T.J.D. and D.R.G. processed Illumina sequence data. D.R.G. processed all ONT data. T.J.D. performed ONT optimization. D.R.G. performed methylation analysis. T.J.D. and D.R.G. performed Illumina and ONT data comparison. C.J.

wrote the original draft. All authors performed manuscript editing.

Acknowledgements

We would like to thank Oxford Nanopore Technologies for supplying the Rapid Barcoding Kit—SQK-RBK001 and FLO-MIN106 R9.4 flow cell used in this research. In particular we would like to thank Leila Luheshi and Divya Mirrington for their support and scientific assistance.

The research was part funded by the National Institute for Health Research Health Protection Research Unit in Gastrointestinal Infections at University of Liverpool in partnership with Public Health England (PHE), in collaboration with University of East Anglia, University of Oxford and the Quadram Institute. Claire Jenkins and Timothy Dallman are based at Public Health England. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR, the Department of Health or Public Health England.

We would also like to thank the frontline NHS Laboratories for submitting the samples used in this study to the Gastrointestinal Bacteria Reference Unit at Public Health England.

We would like to acknowledge Dr Andrew Page of the Quadram Institute for critically reviewing the manuscript.

References

- Croxen MA, Law RJ, Scholz R, et al. Recent advances in understanding enteric pathogenic *Escherichia coli*. *Clin Microbiol Rev* 2013;26(4):822–80.
- Byrne L, Jenkins C, Launders N, et al. The epidemiology, microbiology and clinical impact of Shiga toxin-producing *Escherichia coli* in England, 2009–2012. *Epidemiol Infect* 2015;143(16):3475–87.
- Launders N, Byrne L, Jenkins C, et al. Disease severity of Shiga toxin-producing *E. coli* O157 and factors influencing the development of typical haemolytic uraemic syndrome: a retrospective cohort study, 2009–2012. *BMJ Open* 2016;6(1):e009933.
- Heiman KE, Mody RK, Johnson SD, et al. *Escherichia coli* O157 outbreaks in the United States, 2003–2012. *Emerg Infect Dis* 2015;21(8):1293–301.
- Dallman TJ, Ashton PM, Byrne 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(3):e000029.
- Dallman TJ, Byrne L, Ashton PM, et al. Whole-genome sequencing for national surveillance of Shiga toxin-producing *Escherichia coli* O157. *Clin Infect Dis* 2015;61(3):305–12.
- Olson ND, Lund SP, Colman RE, et al. Best practices for evaluating single nucleotide variant calling methods for microbial genomics. *Front Genet* 2015;6:235.
- Ruffalo M, Koytürk M, Ray S, et al. Accurate estimation of short read mapping quality for next-generation genome sequencing. *Bioinformatics* 2012;28(18):i349–55.
- Timme RE, Rand H, Sanchez Leon M, et al. Genome-Trakr proficiency testing for foodborne pathogen surveillance: an exercise from 2015. *Microb Genom* 2018;4(7), doi:10.1099/mgen.0.000185.
- Quick J, Loman NJ, Duraffour S, et al. Real-time portable genome sequencing for Ebola surveillance. *Nature* 2016;530(7589):228–32.
- Gomez-Eichelmann MC, Levy-Mustri A, Ramirez-Santos J. Presence of 5-methylcytosine in CC(A/T)GG sequences (Dcm methylation) in DNAs from different bacteria. *J Bacteriol* 1991;173(23):7692–4.
- Mikheyev AS, Tin MM. A first look at the Oxford Nanopore MinION sequencer. *Mol Ecol Resour* 2014;14(6):1097–102.
- Laver T, Harrison J, O'Neill PA, et al. Assessing the performance of the Oxford Nanopore Technologies MinION. *Biomol Detect Quantif* 2015;3:1–8.
- Magi A, Semeraro R, Mingrino A, et al. Nanopore sequencing data analysis: state of the art, applications and challenges. *Brief Bioinform* 2018;19(6):1256–72.
- Rang FJ, Kloosterman WP, de Ridder J. From squiggle to base-pair: computational approaches for improving nanopore sequencing read accuracy. *Genome Biol* 2018;19(1):90.
- Senol Cali D, Kim JS, Ghose S, et al. Nanopore sequencing technology and tools for genome assembly: computational analysis of the current state, bottlenecks and future directions. *Brief Bioinform* 2018, doi:10.1093/bib/bby017.
- Marinus MG. DNA methylation in *Escherichia coli*. *Annu Rev Genet* 1987;21:113–31.
- Loman NJ, Quick J, Simpson JT. A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nat Methods* 2015;12(8):733–5.
- Wick RR, Judd LM, Gorrie CL, et al. Completing bacterial genome assemblies with multiplex MinION sequencing. *Microb Genom* 2017;3(10):e000132.
- Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 2014;30(15):2114–20.
- KmerID GitHub repository <https://github.com/phe-bioinformatics/kmerid>.
- Tewolde R, Dallman T, Schaefer U, et al. MOST: a modified MLST typing tool based on short read sequencing. *PeerJ* 2016;4:e2308.
- Gene Finder GitHub repository <https://github.com/phe-bioinformatics/gene.finder>.
- Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 2012;9(4):357–9.
- Li H, Handsaker B, Wysoker A1000 Genome Project Data Processing Subgroup, et al., 1000 Genome Project Data Processing Subgroup The Sequence alignment/map (SAM) format and SAMtools. *Bioinformatics* 2009;25(16):2078–9.
- Ashton PM, Perry N, Ellis R, et al. Insight into Shiga toxin gene encoded by *Escherichia coli* O157 from whole genome sequencing. *PeerJ* 2015;3:e739.
- Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler Transform. *Bioinformatics* 2009;25(14):1754–60.
- McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: a MapReduce frame-work for analyzing next-generation DNA sequencing data. *Genome Res* 2010;20(9):1297–303.
- Dallman T, Ashton P, Schaefer U, et al. SnapperDB: a database solution for routine sequencing analysis of bacterial isolates. *Bioinformatics* 2018;34(17):3028–9.
- De Coster W, D'Hert S, Schultz DT, et al. NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics* 2018;34(15):2666–9.
- Wick R. Porechop GitHub page. <https://github.com/rrwick/Porechop>.
- Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol* 2014;15(3):R46.

33. Page A, Keane J. Rapid multi-locus sequence typing direct from uncorrected long reads using Krocus. *PeerJ* 2018;**6**:e5233.
34. Li H. Minimap2: fast pairwise alignment for long nucleotide sequences. *Bioinformatics* 2018;**34**(18):3094–100.
35. Milne I, Stephen G, Bayer M, et al. Using Tablet for visual exploration of second-generation sequencing data. *Brief Bioinform* 2013;**14**(2):193–202.
36. Croucher NJ, Page AJ, Connor TR, et al. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Res* 2014;**43**(3):e15.
37. Stamatakis A. RAxML Version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 2014;**30**(9):1312–13.
38. Koren S, Walenz BP, Berlin K, et al. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res* 2017;**27**(5):722–36.
39. Hunt M, Silva ND, Otto TD, et al. Circlator: automated circularization of genome assemblies using long sequencing reads. *Genome Biol* 2015;**16**(294):1–10.
40. Wick RR, Judd LM, Gorrie CL, et al. Unicycler: resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Comput Biol* 2017;**13**(6):e1005595.
41. Walker BJ, Abeel T, Shea T, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* 2014;**9**(11):e112963.
42. Bankevich A, Nurk S, Antipov D, et al. SPAdes: a new genome assembly algorithm and its applications to single cell sequencing. *J Comput Biol* 2012;**19**(5):455–77.
43. Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 2014;**30**(14):2068–9.
44. Darling ACE, Mau B, Blattner FR, et al. Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res* 2004;**14**(7):1394–403.
45. Greig DR, Jenkins C, Gharbia S, et al. “Supporting data for “Comparison of single-nucleotide variants identified by Illumina and Oxford Nanopore technologies in the context of a potential outbreak of Shiga toxin-producing *Escherichia coli*.” GigaScience Database 2019; <http://dx.doi.org/10.5524/100623>.

Comparison of Shiga toxin-encoding bacteriophages in highly pathogenic strains of Shiga toxin-producing *Escherichia coli* O157:H7 in the UK

Daniel A. Yara^{1†}, David R. Greig^{2,3†}, David L. Gally³, Timothy J. Dallman^{2,3} and Claire Jenkins^{2,*}

Abstract

Over the last 35 years in the UK, the burden of Shiga toxin-producing *Escherichia coli* (STEC) O157:H7 infection has, during different periods of time, been associated with five different sub-lineages (1983–1995, Ia, I/IIa and I/IIb; 1996–2014, Ic; and 2015–2018, IIb). The acquisition of a *stx2a*-encoding bacteriophage by these five sub-lineages appears to have coincided with their respective emergences. The Oxford Nanopore Technologies (ONT) system was used to sequence, characterize and compare the *stx*-encoding prophages harboured by each sub-lineage to investigate the integration of this key virulence factor. The *stx2a*-encoding prophages from each of the lineages causing clinical disease in the UK were all different, including the two UK sub-lineages (Ia and I/IIa) circulating concurrently and causing severe disease in the early 1980s. Comparisons between the *stx2a*-encoding prophage in sub-lineages I/IIb and IIb revealed similarity to the prophage commonly found to encode *stx2c*, and the same site of bacteriophage integration (*sbcB*) as *stx2c*-encoding prophage. These data suggest independent acquisition of previously unobserved *stx2a*-encoding phage is more likely to have contributed to the emergence of STEC O157:H7 sub-lineages in the UK than intra-UK lineage to lineage phage transmission. In contrast, the *stx2c*-encoding prophage showed a high level of similarity across lineages and time, consistent with the model of *stx2c* being present in the common ancestor to extant STEC O157:H7 and maintained by vertical inheritance in the majority of the population. Studying the nature of the *stx*-encoding bacteriophage contributes to our understanding of the emergence of highly pathogenic strains of STEC O157:H7.

DATA SUMMARY

All FASTQ files and assemblies of samples sequenced in this project have been submitted to the National Center for Biotechnology Information (NCBI) under BioProject accession number PRJNA315192 – <https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA315192>. Strain specific details can be found in Methods under 'Data deposition'. Publicly available data used in this project can be found via Table 1 and in Data Bibliography.

INTRODUCTION

Shiga toxin-producing *Escherichia coli* (STEC) serotype O157:H7 is a zoonotic pathogen that causes gastrointestinal symptoms in humans. A sub-set of patients (mainly children and the elderly) are at risk of developing haemolytic uraemic syndrome (HUS), a potentially fatal systemic condition primarily associated with acute renal failure, and cardiac and neurological complications [1]. STEC O157:H7 emerged as a public-health concern during the early 1980s and was first isolated in the UK in July 1983 from three cases linked to an outbreak of HUS [2]. Throughout the 1980s, the increasing number of outbreaks of gastrointestinal disease, and HUS

Received 15 November 2019; Accepted 17 January 2020; Published 25 February 2020

Author affiliations: ¹Norwich Medical School, University of East Anglia, Norwich, UK; ²National Infection Service, Public Health England, London NW9 5EQ, UK; ³Division of Infection and Immunity, The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Easter Bush EH25 9RG, UK.

*Correspondence: Claire Jenkins, Claire.Jenkins1@phe.gov.uk

Keywords: *Escherichia coli* O157:H7; bacteriophage; Shiga toxin; whole-genome sequencing.

Abbreviations: HUS, haemolytic uraemic syndrome; NCBI, National Center for Biotechnology Information; PHE, Public Health England; PT, phage type; SBI, site of bacteriophage integration; STEC, Shiga toxin-producing *Escherichia coli*; Stx, Shiga toxin.

All FASTQ files and assemblies of samples sequenced in this project have been deposited at the NCBI under BioProject accession number PRJNA315192.

†These authors contributed equally to this work

Data statement: All supporting data, code and protocols have been provided within the article or through supplementary data files.

000334 © 2020 The Authors



This is an open-access article distributed under the terms of the Creative Commons Attribution License.

associated with this serotype, stimulated the development of sub-typing methods that provided a higher level of strain discrimination than serotyping. In the late 1980s, a phage typing scheme, developed by the Canadian Public Health Laboratory Service, was adopted by Public Health England (PHE; then the Public Health Laboratory Service) [3], and is still used today. In 2015, PHE implemented whole-genome sequencing for routine surveillance of STEC O157:H7 in England [4].

The primary STEC virulence factor is the Shiga toxin (Stx), which targets cells expressing the glycolipid globotriaosylceramide, disrupting host protein synthesis and causing apoptotic cell death. Strains of STEC O157:H7 in the UK produce *stx1a*, *stx2a* and *stx2c*, either individually or in any combination [5]. Strains harbouring *stx2a*, either alone or in combination with *stx1a* and/or *stx2c*, are significantly associated with causing severe disease, including HUS [5, 6], and are associated with more efficient transmission within the ruminant reservoir [7]. The genes encoding the *stx* subtypes are located on active bacteriophage that can be acquired and integrated into the chromosome of STEC O157:H7 strains. There is evidence that the different prophage backgrounds that harbour *stx* genes can contribute to differential toxin production and may ultimately affect clinical outcome [8].

There are three main lineages of STEC O157:H7 (I, II and I/II) and eight sub-lineages (Ia, Ib, Ic, IIa, IIb, IIc, I/IIa and I/IIb). In the UK, the outbreaks of STEC O157:H7 in the 1980s were caused by strains belonging to sub-lineage Ia [mainly comprising phage type (PT)1 and PT4], sub-lineage I/IIa (comprising PT2) and sub-lineage I/IIb (comprising PT49) [9]. Throughout the 1990s, these three lineages declined and almost disappeared. Concurrently, we observed a dramatic rise of sub-lineage Ic (mainly comprising PT21/28), in addition to a steady increase in the number of cases of sub-lineage IIc (mainly comprising PT8) [5, 9]. Since 2012, the number of cases of PT21/28 has declined and an unusual PT8 variant belonging to sub-lineage IIb has emerged [10].

With the exception of sub-lineage IIc (PT8), which is not associated with HUS cases in the UK [5], all the dominant UK sub-lineages over time encode *stx2a*, and the acquisition of a *stx2a*-encoding bacteriophage appears to have coincided with their respective emergences [5, 10]. The aim of this investigation was to use the Oxford Nanopore Technologies system to sequence, characterize and compare the *stx*-encoding prophage harboured by each of the UK sub-lineages to determine the similarity of the *stx*-encoding prophage acquired by each lineage. Studying the nature of the *stx*-encoding bacteriophage will contribute to our understanding of the emergence of highly pathogenic strains of STEC O157:H7.

METHODS

Bacterial strains

Six strains of STEC O157:H7 were selected for sequencing from the PHE archive on the basis of being the earliest representative of each of the sub-lineages that acquired the *stx2a*-encoding prophage (Table 1, Fig. 1). Eleven publicly available sequences

Impact Statement

The application of the Oxford Nanopore Technologies system to sequence UK epidemic strains of Shiga toxin-producing *Escherichia coli* (STEC) O157:H7 revealed *stx2a*-encoding prophages exhibit a high level of diversity. There was little evidence of geographical or temporal patterns of relatedness, or of intra-UK transmission of *stx2a*-encoding prophage between indigenous strains. The *stx2a*-encoding prophages in the UK lineages associated with severe disease appear to be acquired independently and most likely from different geographical and/or environmental sources. These data provide supporting evidence for the existence of a dynamic environmental reservoir of *stx2a*-encoding prophages that pose a threat to public health due to their potential for integration into competent, indigenous sub-lineages of STEC O157:H7. We also provide further evidence that *stx2c*-encoding prophages exhibit a high level of similarity across lineages, geographical regions and time, and have likely been maintained and inherited vertically.

were also included in the analysis for context. Of these, seven originated from the UK, five were the cause of four published outbreaks [11–13], three were from the USA [14, 15] and one was from Japan [16] (Table 1, Fig. 1).

Short-read sequencing on the Illumina HiSeq 2500

Genomic DNA was extracted from cultures of STEC O157:H7 using the QIAasympy system (Qiagen). The sequencing library was prepared using the Nextera XP kit (Illumina) for sequencing on the HiSeq 2500 instrument (Illumina), run with the fast protocol. FASTQ reads were processed using Trimmomatic v0.27 [17] to remove bases with a PHRED score of <30 from the leading and trailing ends, with reads <50 bp after quality trimming discarded.

Long-read sequencing and data processing

Genomic DNA was extracted and purified using the Qiagen genomic tip, midi 100/G, with minor alterations including no vigorous mixing steps (mixing performed by inversion instead) and elution into 100 µl double processed nuclease-free water (Sigma-Aldrich). Genomic DNA for each extract was quantified using a Qubit and the HS (high sensitivity) dsDNA assay kit (ThermoFisher Scientific), following the manufacturer's instructions. Library preparation was performed for several instances using both rapid barcoding [SQK-RBK00(1/4)] and native barcoding kits (SQK-LSK108 and EXP-NBD103) (Oxford Nanopore Technologies). The prepared libraries were loaded onto FLO-MIN106 R9.4.1 flow cells (Oxford Nanopore Technologies) and sequenced using the MinION (Oxford Nanopore Technologies) for 48 h.

Data produced in a raw FAST5 format was basecalled and de-multiplexed using Albacore v2.3.3 (Oxford Nanopore

Table 1. Summary of the PHE archived and publicly available strains used within this study, with their strain ID, lineage, PT, Stx profile, chromosome size (bp), number of prophages in the chromosome, *stx*-encoding prophages with SBI and size (bp), assembly accession numbers and NCBI BioProject accession numbers

Strain ID	Lineage	Phage type	Stx profile	Chromosome size (bp)	No. of prophages	<i>stx1a</i> prophage size (bp)	<i>stx2a</i> prophage SBI and size (bp)	<i>stx2c</i> prophage SBI and size (bp)	Reference	BioProject accession no.	Assembly accession no.
PHE archive											
E30228	Ia	PT4	Stx1a/2a	5 416 109	15	<i>yehV</i> (47 594)	<i>wrbA</i> (62 890)	–	[41]	PRJNA315192	VXJO00000000
E34500	I/Ia	PT2	Stx2a/2c	5 359 964	14	–	<i>argW</i> (62 149)	<i>shcB</i> (57 463)	[2]	PRJNA315192	VXJN00000000
E45000	I/Ib	PT49	Stx2a	5 386 698	17	–	<i>shcB</i> (44 014)	–	This study	PRJNA315192	VXJN00000000
E116508	Ic	PT21/28	Stx2a/2c	5 571 891	17	–	<i>argW</i> (71 870)	<i>shcB</i> (59 105)	This study	PRJNA315192	VXJP00000000
315176	I/Ib	PT8	Stx2a	5 579 120	16	–	<i>shcB</i> (61 851)	–	[10]	PRJNA315192	VXJO00000000
267849	I/Ia	PT34	Stx2a/2c	5 510 912	16	–	<i>yaeE</i> (47 242)	<i>shcB</i> (61 840)	[43]	PRJNA315192	VXJR00000000
Publicly available											
9000	Ic	PT21/28	Stx2a/2c	5 516 497	17	–	<i>argW</i> (65 158)	<i>shcB</i> (57 408)	[33]	PRJNA336330	CP018252
397404	Ic	PT21/28	Stx2a/2c	5 618 435	13	–	<i>argW</i> (70 472)	<i>shcB</i> (59 098)		PRJNA315192	CP043019
155	Ic	PT32	Stx2a	5 513 008	18	–	<i>yaeE</i> (50 015)	–	[33]	PRJNA336330	CP018237
350	I/Ic	PT8	Stx1a/2c	5 411 823	16	<i>yehV</i> (49 867)	–	<i>shcB</i> (57 747)	[11]	PRJNA336330	CP018243
272	I/Ia	PT2	Stx2a	5 474 193	16	–	<i>argW</i> (65 675)	–	[13]	PRJNA336330	CP018239
644	I/Ic	PT8	Stx1a/1a/2c	5 831 209	18	<i>yehV</i> (49 544) <i>argW</i> (64 569)	–	<i>shcB</i> (58 210)	[12]	PRJNA321984	CP015831
180	I/Ic	PT54	Stx1a/1a/2c	5 509 528	15	<i>yehV</i> (49 544) <i>argW</i> (64 569)	–	<i>shcB</i> (61 558)	[12]	PRJNA321984	CP015832
Sakai	Ia	–	Stx1a/2a	5 498 450	18	<i>yehV</i> (47 650)	<i>wrbA</i> (62 142)	–	[16]	PRJNA57781	NC_002695
EDL933	Ia	–	Stx1a/2a	5 547 323	14	<i>yehV</i> (47 596)	<i>wrbA</i> (61 066)	–	[14]	PRJNA253471	CP008957
EC4115	I/Ia	–	Stx2a/2c	5 572 075	17	–	<i>argW</i> (71 540)	<i>shcB</i> (60 476)	[15]	PRJNA224116	NC_011353
TW14359	I/Ia	–	Stx2a/2c	5 528 136	17	–	<i>argW</i> (71 540)	<i>shcB</i> (60 476)	[15]	PRJNA224116	NC_013008

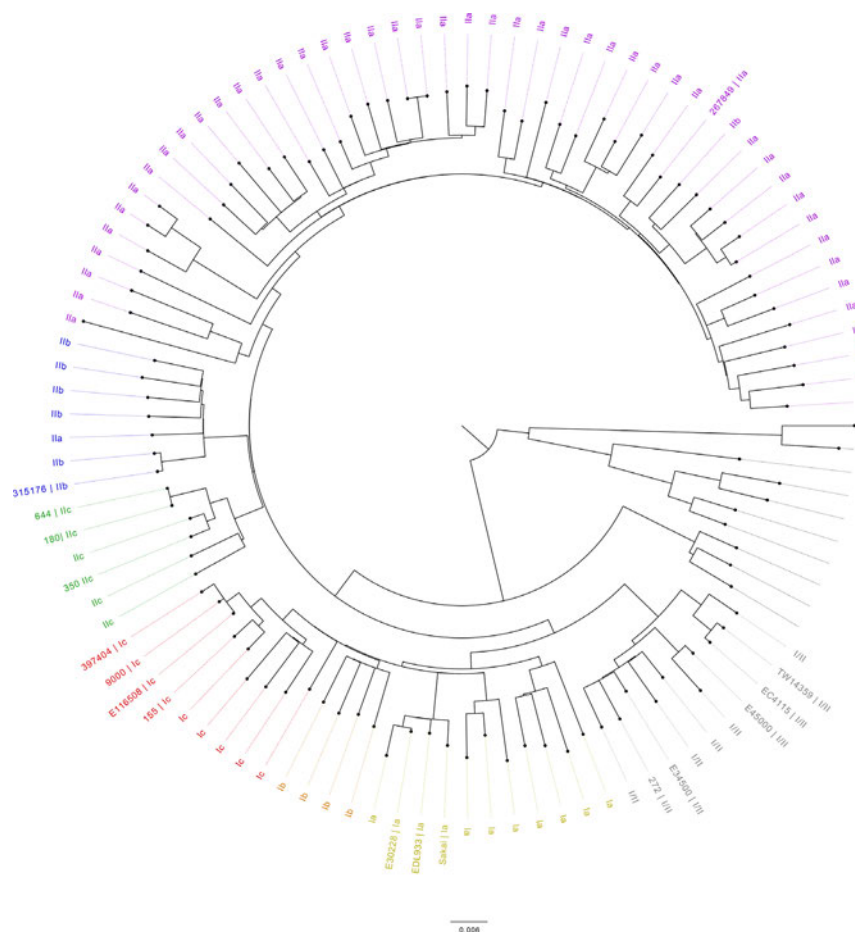


Fig. 1. Maximum-likelihood phylogenetic tree of 105 genomes including the 17 (labelled) publicly available genomes and nanopore sequenced genomes produced during this study. Sub-lineages are coloured as follows: I/II, grey; Ia, yellow; Ib, orange; Ic, red; IIa, purple; IIb, blue; IIc, green. Scale bar indicates kbp.

Technologies) into FASTQ format and grouped in each samples' respective barcode. De-multiplexing was performed using Deepbiner v0.2.0 [18]. Run metrics were generated using Nanoplot v1.8.1 [19]. The barcode and y-adaptor from each sample's reads were trimmed, and chimeric reads split using Porechop v0.2.4 [20]. Finally, the trimmed reads were filtered using Filtlong v0.1.1 [21] with the following parameters, min length=1000 bp, keep per cent=90 and target bases=550 Mbp, to generate approximately 100× coverage of the STEC genome with the longest and highest-quality reads.

De novo assembly, polishing, reorientation and annotation

Trimmed nanopore FASTQ files were assembled using Canu v1.7 [22] and the filtered nanopore FASTQ files were assembled using both Unicycler v0.4.2 [23], with the following parameters min_fasta_length=1000 bp, mode=normal, and Flye v2.4.2 [24], using default parameters. The assembly for each sample that had the highest N50 and lowest number of contigs with the assembly size (between 5.3–6.0 Mbp) were taken forward. Polishing of the assemblies was performed in a three-step process. Firstly, polishing was initiated using

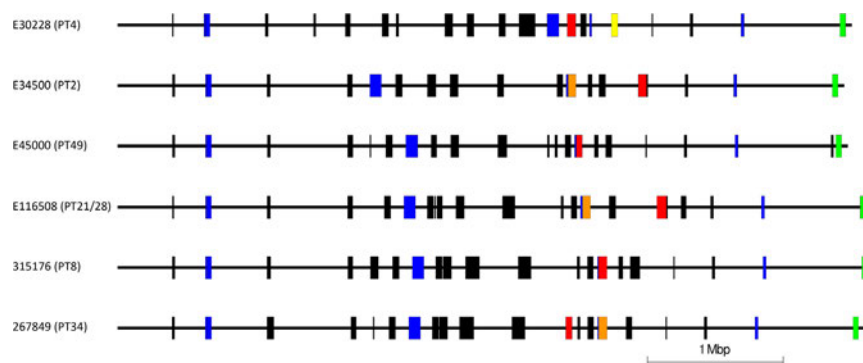


Fig. 2. Easyfig diagram representing the chromosome and prophage content within the samples sequenced in this study (in descending order PT4, PT2, PT49, PT21/28, PT8 and PT34). *stx2a*-encoding, *stx2c*-encoding and *stx1*-encoding prophages are highlighted in red, orange and yellow, respectively. Non-*stx*-encoding prophages are coloured black. Prophage-like elements are coloured blue and the locus of enterocyte effacement is shown in green.

Nanopolish v0.11.1 [25] using both the trimmed nanopore FASTQs and FAST5s for each respective sample accounting for methylation using the `--methylation-aware=dc` and `--min-candidate-frequency=0.5`. Secondly, the polishing was continued with Pilon v1.22 [26] using Illumina FASTQ reads as the query dataset with the use of BWA v0.7.17 [27] and Samtools v1.7 [28]. Finally, Racon v1.2.1 [29] also using BWA v0.7.17 [27] and Samtools v1.7 [28] was used with the Illumina reads for two cycles to produce a final assembly for each of the samples. As the chromosome from each assembly was circularized and closed, they were re-orientated to start at the *dnaA* gene (GenBank accession no. NC_000913) from *E. coli* K12, using the `--fixstart` parameter in Circlator v1.5.5 [30]. Prokka v1.13 [31] with the use of a personalized database (an amino acid FASTA that included all genes annotated in the publicly available samples used in this study) was used to annotate the final assemblies.

Prophage detection, excision and processing

Prophages across all samples were detected and extracted using the updated Phage Search Tool (PHASTER) [32]. Prophage extraction from the genome occurred regardless of prophage size or PHASTER quality score, and any detected prophages separated by less than 4 kbp were conjoined into a single phage using Propi v0.9.0, as described elsewhere [33]. From here, the prophages were trimmed to remove any non-prophage genes and were again annotated using Prokka v1.13 [31] with the use of a personalized database (an amino acid FASTA that included all genes annotated in the publicly available samples used in this study).

Mash and Stx-encoding prophage phylogeny

Mash v2.2 [34] was used to sketch (sketch length 1000 bp, kmer length 21) the extracted prophages in the samples sequenced in this study and all Stx-encoding prophages found

in the publicly available STEC genomes in Table 1. The pairwise Jaccard distance between the prophages was calculated and a neighbour-joining tree computed and visualized using FigTree v1.4.4 [35].

Visualization tools and phylogenetic context

To provide context for the 17 nanopore sequenced ($n=6$) and publicly available ($n=11$) samples, a maximum-likelihood tree was recreated combining these 17 genomes with 88 genomes (105 in total) from PHE's STEC collection (clonal complex 11). Each of the 88 selected genomes is from a unique 250 single-linkage hierarchical cluster that was calculated using SnapperDB v0.2.6 [36]. SnapperDB was also used to generate a whole-genome alignment of all 105 genomes that was processed through Gubbins v2.00 [37] to identify any recombinant sequences. The tree was recreated by RAxML v8.2.8 [38]. Phylogenetic trees were visualized and annotated using FigTree v1.4.4 [35]. All gene diagrams were constructed using Easyfig v2.2.3 [39].

Data deposition

Illumina FASTQ files are available from National Center for Biotechnology Information (NCBI) BioProject PRJNA315192 under the following SRA (sequence read archive) accession numbers: E30228, SRR10290290; E34500, SRR10290289; E45000, SRR10290288; E116508, SRS941727; 315176, SRR6051955; and 267849, SRR3742262. Nanopore FASTQ files are available from BioProject PRJNA315192 under the following SRA accession numbers: E30228, SRR10103064; E34500, SRR10103063; E45000, SRR10103062; E116508, SRR10103065; 315176, SRR10103066; and 267849, SRR10103067. Assemblies can be found under BioProject PRJNA315192 under the following accession numbers: E30228, VXJO000000000; E34500, VXJN000000000; E45000, VXJM000000000; E116508, VXJP000000000; 315176, VXJQ000000000; and 267849, VXJR000000000.

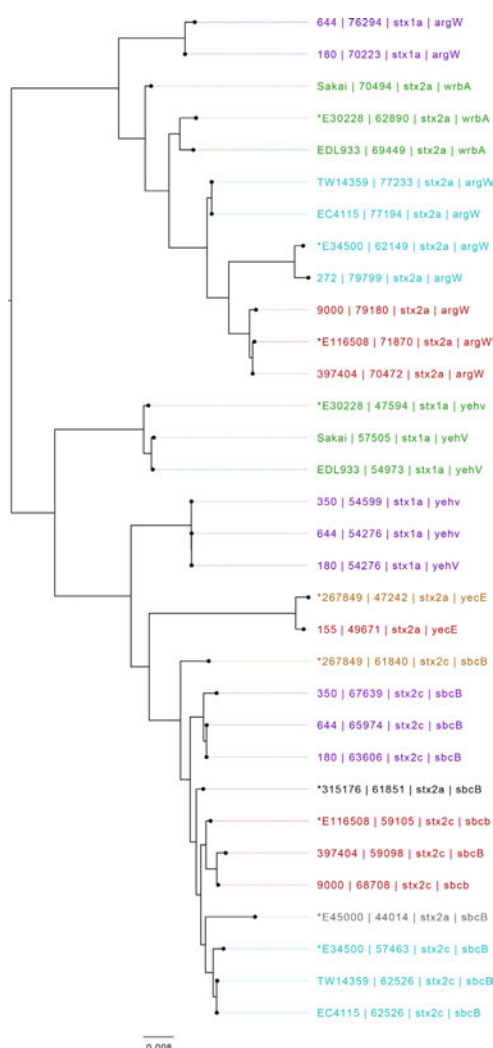


Fig. 3. Mid-rooted tree of *stx*-encoding prophages based on Jaccard distance produced from Mash. Strains are annotated with strain ID, length (bp), *stx* profile and SBI. Strains sequenced during this study have prophages that are hown preceded by an *. Strains are coloured by sub-lineage: green, Ia; red, Ic; blue, I/IIa; grey, I/IIb; orange, IIa; black, IIb; purple, IIc. Scale bar indicates Jaccard distance.

RESULTS AND DISCUSSION

Genomic features of the samples sequenced in this study

All six isolates, selected for sequencing from the PHE archive

on the basis of being the earliest representative of each of the sub-lineages that acquired the *stx2a*-encoding prophage, assembled into closed chromosomes with one or more plasmids. The isolates belonging to sub-lineage Ia PT4 (E30228) and sub-lineage IIb PT8 (315176) each assembled into a chromosome (5 416 109 and 5 579 120bp, respectively) and two plasmids (Table 1). The sequence data from the other four isolates each assembled into a chromosome of between 5 359 964 and 5 571 891 bp and a single plasmid (Table 1). The pO157 (IncFIB) plasmid was found in all samples sequenced in this study. The number of prophages in each of the genomes of the six isolates varied from 14 in the isolate belonging to sub-lineage I/IIa PT2 to 17 from the isolates belonging to sub-lineages I/IIb PT49 and Ic PT21/28 (Fig. 2).

Comparison of the *stx1a*-encoding prophage

Six of the isolates analysed in this study contained a prophage encoding *stx1a* (Table 1, Figs 3 and 4). The *stx1a*-encoding prophage from the isolate belonging to sub-lineage Ia PT4 (E30228), among the first to be isolated in the UK in 1983, shared similarity with *stx1a*-encoding prophage found in EDL933 and Sakai, two international outbreak strains that also belong to sub-lineage Ia (Table 1, Figs 3 and 4). EDL933 caused an outbreak in the USA in 1982 linked to contaminated hamburgers [14], and was temporally but not geographically linked to the UK isolate. The outbreak in Sakai City, Japan, associated with contaminated radish sprouts, occurred in 1996 [16], and was both temporally and geographically distinct from EDL933 and E30228 (Figs 3 and 4). Previous analysis of isolates of sub-lineage Ia harbouring *stx1a*-encoding prophage indicate the *stx1a* prophage is likely ancestral and inherited vertically [5]. This is consistent with the strains analysed in this study encoding a similar *stx1a* prophage, despite being isolated at different times and geographical locations.

The *stx1a*-encoding prophages from three isolates belonging to sub-lineage IIc associated with foodborne outbreaks in the UK [11, 12] cluster together based on Mash distance, but were distinct from the *stx1a*-encoding prophages harboured by the sub-lineage Ia strains described above. As previously described [33, 40], two of these strains (664 PT8 and 180 PT54), linked to a foodborne outbreak in Northern Ireland in 2013 [12], had an additional but different *stx1a*-encoding prophage within the same chromosome (Fig. 4). Therefore, three different *stx1a*-encoding prophages, in two different lineages (Ia and IIc), were identified in this study (Figs 3 and 4).

Comparison of *stx2c*-encoding prophage

Nine isolates from four different sub-lineages (Ic, I/IIa, IIa and IIc) contained *stx2c*-encoding prophage. The *stx2c*-encoding prophage from each of the isolates clustered together based on Mash distance and also aligned across the length of the prophage with few structural variations (Fig. 5). The *stx2c* prophage from strains within the same sub-lineage were more similar based on Mash distance than *stx2c* prophage in strains from different lineages (Table 1, Figs 3 and 5). These strains were isolated over a wide time frame from 1983 to 2016, and

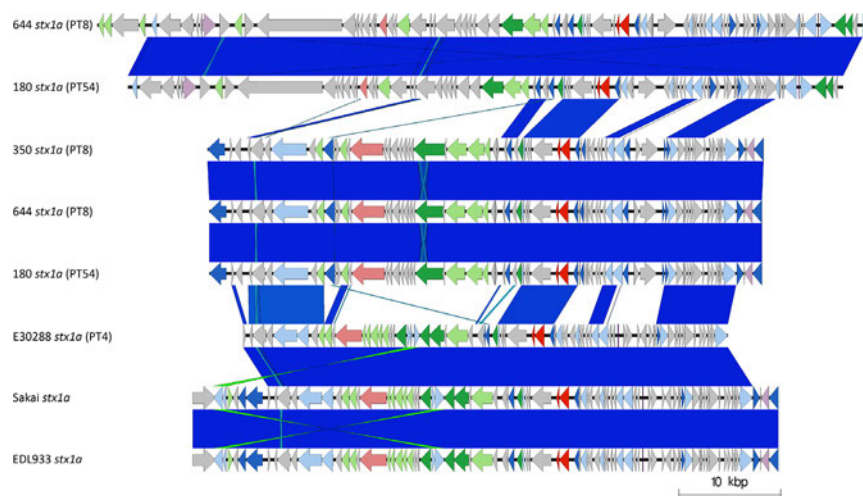


Fig. 4. Easyfig plot comparing the *stx1a*-encoding prophages from 644 (×2), 180 (×2), 350, E30288, EDL933 and Sakai. Arrows indicate gene directions. *stx* genes are shown in red; recombination/replication genes are shown in light blue; regulation-associated genes are shown in dark blue; effector genes are shown in pink; structure- and lysis-associated genes are shown in light and dark green, respectively; tRNAs are shown as purple lines; finally, hypothetical genes are shown in grey.

in different countries including the UK, Ireland and the USA, providing further evidence that *stx2c*-encoding prophages show a high level of similarity across lineages, time and geographical regions [33] (Table 1). This is consistent with the model of *stx2c* being present in the common ancestor to extant STEC O157:H7 and maintained by vertical inheritance in the majority of the population.

Comparison of *stx2a*-encoding prophage

Certain strains that shared lineage, PT and geography harboured similar *stx2a*-encoding prophages. Examples included (i) the two sub-lineage Ic PT21/28 isolates from the UK, (ii) the two sub-lineage I/IIa PT2 isolates from the UK and (iii) the two isolates from sub-lineage I/IIa from the USA (Table 1, Figs 3 and 6). Isolates designated E30228 and EDL933, both sub-lineage Ia and temporally related but geographically distinct, also encoded similar *stx2a*-encoding prophage (Table 1, Figs 3 and 6), as did isolates 155 (sub-lineage Ic PT32) and 267849 (sub-lineage IIa PT34), which were unrelated temporally and geographically.

Compared to *stx2c* prophage, however, the *stx2a*-encoding prophage found in 11 of the isolates in this study exhibited a greater diversity both based on Mash distance and whole-prophage alignment. The *stx2a*-encoding prophage from each of the lineages causing severe clinical disease in the UK were all distinct, including the two UK sub-lineages (Ia and I/IIa) circulating concurrently and causing outbreaks of HUS in the early 1980s [2, 41] (Fig. 3). Throughout the 1980s, the number of sub-lineage Ia strains (mainly PT1 and PT4) declined and

a new sub-lineage, I/IIb PT49, emerged. The *stx2a* in the emerging sub-lineage I/IIb PT49 strain was encoded on a bacteriophage that was again distinct from either of the two *stx2a*-encoding prophages found in the representative isolates from the early contemporary sub-lineages Ia and I/IIa. Comparisons between the *stx2a*-encoding prophages in sub-lineage I/IIb revealed similarity to the prophages commonly found to encode *stx2c* (Figs 3 and 5). Furthermore, sub-lineage I/IIb *stx2a*-encoding prophages had the same site of bacteriophage integration (SBI) as sub-lineage I/IIa *stx2c*-encoding prophages, specifically the *sbcB* gene.

During the 1990s, all three of the dominant 1980s sub-lineages (Ia, I/IIa and I/IIb) declined as a cause of human gastrointestinal disease, and a new sub-lineage emerged. STEC O157:H7 *stx2c* PT32 belonging to sub-lineage Ic had been circulating in UK and Irish cattle populations for many decades, but had not been linked to cases of human disease [5]. However, following acquisition of a *stx2a*-encoding prophage (into the SBI *argW*), which resulted in a change in PT to PT21/28 [5, 33], sub-lineage Ic became the most common STEC O157:H7 sub-lineage causing gastrointestinal disease and HUS in humans in the UK for the next two decades. The *stx2c*-encoding prophage in lineage Ic had high sequence similarity to *stx2c*-encoding prophages in the other isolates analysed in this study and shared the same SBI, *sbcB* (Table 1, Figs 2 and 5). However, the *stx2a*-encoding prophage acquired by sub-lineage Ic once again differed from those found in the three sub-lineages circulating in the previous decade (Table 1, Figs 2 and 6).

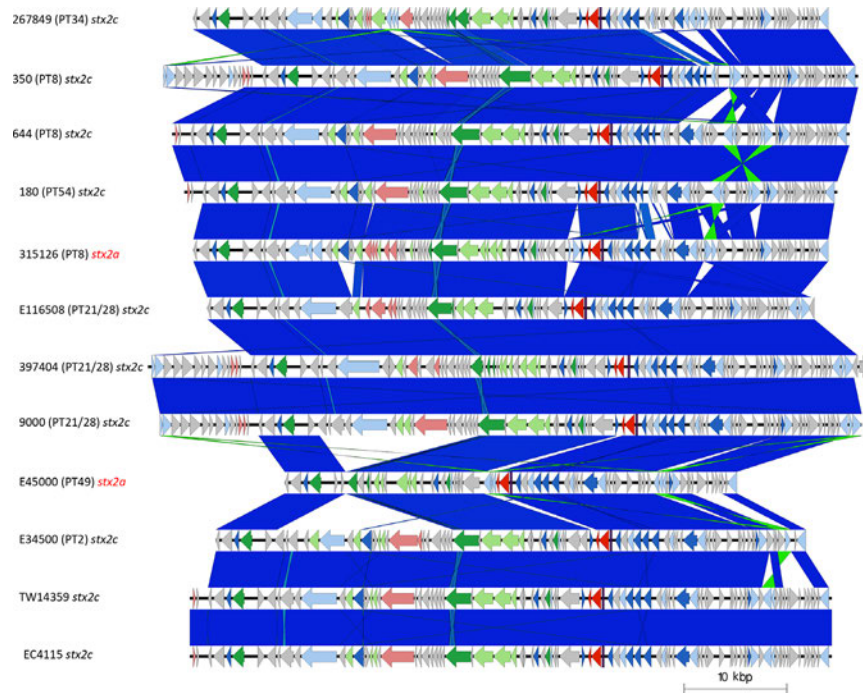


Fig. 5. Easyfig plot comparing the *stx2c*-encoding prophages from all samples in the study, including two *stx2a* prophages that are in a *stx2c*-associated prophage structure (315126 and E45000). Arrows indicate gene directions. *stx* genes are shown in red; recombination/replication genes are shown in light blue; regulation-associated genes are shown in dark blue; effector genes are shown in pink; structure- and lysis-associated genes are shown in light and dark green, respectively; tRNAs are shown as purple lines; finally, hypothetical genes are shown in grey.

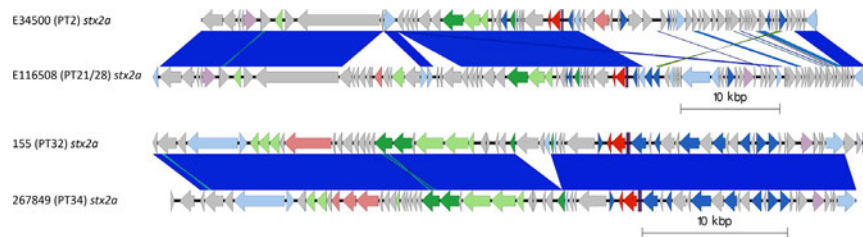


Fig. 6. Two Easyfig plots comparing the *stx2a*-encoding prophages from E45000 with E116508 (above) and 155 and 267849 (below), in descending order. Arrows indicate gene directions. *stx* genes are shown in red; recombination/replication genes shown in light blue; regulation-associated genes are shown in dark blue; effector genes are shown in pink; structure- and lysis-associated genes are shown in light and dark green, respectively; tRNAs are shown as purple lines; finally, hypothetical genes are shown in grey.

Recently, in the UK, there has been a decrease in the number of cases caused by STEC O157:H7 belonging to sub-lineage Ic, and an emergence of sub-lineage IIb PT8 that appears to be associated with the acquisition of a prophage encoding *stx2a* [9]. Strains belonging to this sub-lineage have caused food-borne outbreaks linked to contaminated mixed-leaf salad, lamb-based meat products including sausages and mince [42], and an environmental exposure linked to participation in a mud-based obstacle event [42]. Like the *stx2a*-encoding prophage described in sub-lineage I/IIb, the *stx2a*-encoding prophage in sub-lineage IIb was similar to the *stx2c*-encoding prophage, and likely the result of horizontal exchange of the *stx2a* gene into a previously *stx2c*-encoding prophage. This is also corroborated by the *stx2a*-encoding prophage in sub-lineage IIb integrating at *sbvB* associated with *stx2c*-encoding prophages (Table 1, Figs 3, 5 and 6).

Importation of STEC O157:H7 strains from outside the UK via contaminated food products is a constant threat. In 2016, a large national outbreak of STEC O157:H7 *stx2a/stx2c* PT34 belonging to sub-lineage IIa occurred in the UK [43]. Epidemiological investigations concluded that contaminated red Batavia salad leaves from a non-domestic source was the most plausible vehicle of infection. Analysis of the nanopore data from the outbreak strain demonstrated that the *stx2a*-encoding prophage was different from all the *stx2a*-encoding prophages identified in the five major UK sub-lineages. However, this prophage shared sequence similarity with the *stx2a*-encoding prophage in STEC O157:H7 PT32 belonging to sub-lineage Ic, associated with cases of severe gastrointestinal disease in Ireland [5] (Fig. 6). Unlike the previously described *stx2a*-encoding prophage, the *stx2a*-encoding prophage in both of these strains share the SBI *yecE*. This prophage also had similarity to the *stx2a*-encoding prophage found in a strain of STEC O55:H7 causing recurrent, seasonal outbreaks of HUS in England [40].

Summary

Currently, the application of nanopore technology for extensive characterization of STEC O157:H7 genomes at PHE is still under development; therefore, the number of sequences analysed in this study was limited. *stx2a*-encoding prophages exhibited a higher level of diversity and there was little evidence of geographical or temporal patterns of relatedness, or of intra-UK transmission of *stx2a*-encoding prophage between indigenous strains. The *stx2a*-encoding prophages in the UK lineages associated with severe disease, therefore, appear to be acquired independently and most likely from different geographical and/or environmental sources. These data provide supporting evidence for the existence of a dynamic environmental reservoir of *stx2a*-encoding prophages that pose a threat to public health due to their potential for integration into competent, indigenous sub-lineages of *E. coli* O157:H7. Finally, we provide further evidence that, compared to *stx2a*-encoding prophages, *stx2c*-encoding prophages exhibit a high level of similarity across lineages, geographical regions and time, and have likely been maintained and inherited vertically.

Funding information

The research was part funded by the National Institute for Health Research (NIHR) Health Protection Research Unit in Gastrointestinal Infections at the University of Liverpool (UK), in partnership with PHE, in collaboration with the University of East Anglia (UK), the University of Oxford (UK) and the Quadram Institute (UK). C.J., T.J.D. and D.R.G. are based at PHE. The views expressed are those of the authors and not necessarily those of the National Health Service, the NIHR, the Department of Health nor PHE. D.A.Y. was funded by a Doctoral Training Partnership PhD studentship from the BBSRC (Biotechnology and Biological Sciences Research Council, UK; <https://bbsrc.ukri.org/>). The funders had no role in study design, data collection and analysis, decision to publish nor preparation of the manuscript.

Author contributions

T. J. D. and C. J. conceptualized the project. D. R. G. performed DNA extractions, library preparations and sequencing of isolates. D. A. Y. and D. R. G. performed data processing, genome assembly, genome polishing and genome annotation. D. R. G. and D. A. Y. created the Easyfig diagrams. D. R. G. performed prophage comparison using Mash and T. J. D. wrote associated scripts. D. A. Y., D. R. G., T. J. D. and C. J. wrote the original manuscript. D. A. Y., D. R. G., T. J. D., C. J. and D. L. G. reviewed and edited the manuscript. T. J. D., C. J. and D. L. G. supervised D. R. G.; whilst D. R. G., T. J. D. and C. J. supervised D. A. Y.

Conflicts of interest

The authors declare that there are no conflicts of interest.

Data Bibliography

1. Yara DA, Greig DR, Gharbia SE, Gally DL, Dallman TJ, Jenkins C. BioProject: PRJNA315192, VXJ000000000-VXJR000000000 (2019).
2. Greig DR, Jenkins C, Gally DL, Gharbia SE, Dallman TJ. BioProject: PRJNA315192-CP043019 (2019).
3. Cowley LA, Dallman TJ, Fitzgerald S, Irvine N, Rooney PJ, McAteer SP, Day M, Perry NT, Bono JL, Jenkins C, Gally DL. GenBank accession numbers CP015831 and CP015832 (2016).
4. Dallman TJ, Byrne L, Ashton PM, Cowley LA, Perry NT, Adak G, Petrovskaya L, Ellis RJ, Elson R, Underwood A, Green J, Hanage WP, Jenkins C, Grant K, Wain J. BioProject PRJNA248042 (2015).
5. Eppinger M, Sebastian Y, Ravel J. NCBI accession number NC_011353 (2016).
6. Latif H, Aziz RK, Charusanti P, Palsson BO. GenBank accession number CP008957 (2014).
7. Makino K, Yokoyama K, Kubota Y, Yutsudo CH, Kimura S, Kurokawa K, Ishii K, Hattori M, Tatsuno I, Abe H, Iida T, Yamamoto K, Onishi M, Hayashi T, Yasunaga T, Honda T, Sasakawa C, Shinagawa H. NCBI accession number NC_002695 (2016).
8. Brittnacher M, Jacobs M, Zhou Y, Chang J, Fong C, Gillett W, Haugen E, Hayden H, Kulasekara B, Larson Freeman T, Radey M, Rohmer L, Sims E, Wu Z, Whittam T, Kaul R, Olson MV, Miller SI. NCBI accession number NC_013008 (2016).
9. Shaaban S, Cowley L, McAteer SP, Jenkins C, Dallman TJ, Bono JL, Gally DL. GenBank accession numbers CP018252 and CP018237 (2016).
10. Launders N, Locking ME, Hanson M, Willshaw G, Charlett A, Salmon R, Cowden J, Harker KS, Adak GK. GenBank accession numbers CP018243 and CP018239 (2016).

References

1. Tarr PI, Gordon CA, Chandler WL. Shiga-toxin-producing *Escherichia coli* and haemolytic uraemic syndrome. *The Lancet* 2005;365:1073–1086.
2. Taylor CM, White RH, Winterborn MH, Rowe B. Haemolytic-uraemic syndrome: clinical experience of an outbreak in the West Midlands. *BMJ* 1986;292:1513–1516.
3. Khakhria R, Duck D, Lior H. Extended phage-typing scheme for *Escherichia coli* O157:H7. *Epidemiol Infect* 1990;105:511–520.
4. 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.

5. 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.
6. Persson S, Olsen KEP, Ethelberg S, Scheutz F. Subtyping method for *Escherichia coli* Shiga toxin (verocytotoxin) 2 variants and correlations to clinical manifestations. *J Clin Microbiol* 2007;45:2020–2024.
7. Fitzgerald SF, Beckett AE, Palarea-Albaladejo J, McAteer S, Shaaban S et al. Shiga toxin sub-type 2a increases the efficiency of *Escherichia coli* O157 transmission between animals and restricts epithelial regeneration in bovine enteroids. *PLoS Pathog* 2019;15:e1008003.
8. Ogura Y, Mondal SI, Islam MR, Mako T, Arisawa K et al. The Shiga toxin 2 production level in enterohemorrhagic *Escherichia coli* O157:H7 is correlated with the subtypes of toxin-encoding phage. *Sci Rep* 2015;16:16663.
9. Adams NL, Byrne L, Smith GA, Elson R, Harris JP et al. Shiga toxin-producing *Escherichia coli* O157, England and Wales, 1983–2012. *Emerg Infect Dis* 2016;22:590–597.
10. Byrne L, Dallman TJ, Adams N, Mikhail AFW, McCarthy N et al. Highly pathogenic clone of Shiga toxin-producing *Escherichia coli* O157:H7, England and Wales. *Emerg Infect Dis* 2018;24:2303–2308.
11. Launders N, Locking ME, Hanson M, Willshaw G, Charlett A et al. A large great Britain-wide outbreak of STEC O157 phage type 8 linked to handling of raw leeks and potatoes. *Epidemiol Infect* 2016;144:171–181.
12. Cowley LA, Dallman TJ, Fitzgerald S, Irvine N, Rooney PJ et al. Short-term evolution of Shiga toxin-producing *Escherichia coli* O157:H7 between two food-borne outbreaks. *Microb Genom* 2016;2:e000084.
13. Jenkins C, Dallman TJ, Launders N, Willis C, Byrne L et al. Public health investigation of two outbreaks of Shiga toxin-producing *Escherichia coli* O157 associated with consumption of watercress. *Appl Environ Microbiol* 2015;81:3946–3952.
14. Riley LW, Remis RS, Helgerson SD, McGee HB, Wells JG et al. Hemorrhagic colitis associated with a rare *Escherichia coli* serotype. *N Engl J Med* 1983;308:681–685.
15. Uhlich GA, Sinclair JR, Warren NG, Chmielecki WA, Fratamico P. Characterization of Shiga toxin-producing *Escherichia coli* isolates associated with two multistate food-borne outbreaks that occurred in 2006. *Appl Environ Microbiol* 2008;74:1268–1272.
16. Michino H, Araki K, Minami S, Takaya S, Sakai N et al. Massive outbreak of *Escherichia coli* O157: H7 infection in schoolchildren in Sakai City, Japan, associated with consumption of white radish sprouts. *Am J Epidemiol* 1999;150:787–796.
17. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 2014;30:2114–2120.
18. Wick RR, Judd LM, Holt KE. Deepbinner: demultiplexing barcoded Oxford Nanopore reads with deep convolutional neural networks. *PLoS Comput Biol* 2018;14:e1006583.
19. De Coster W, D'Hert S, Schultz DT, Cruts M, Van Broeckhoven C. NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics* 2018;34:2666–2669.
20. Wick RR. Porechop; 2017. <https://github.com/rrwick/Porechop>
21. Wick RR. Filtlong; 2017. <https://github.com/rrwick/Filtlong>
22. 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.
23. Wick RR, Judd LM, Gorrie CL, Holt KE. Unicycler: resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Comput Biol* 2017;13:e1005595.
24. Kolmogorov M, Yuan J, Lin Y, Pevzner PA. Assembly of long, error-prone reads using repeat graphs. *Nat Biotechnol* 2019;37:540–546.
25. Loman NJ, Quick J, Simpson JT. A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nat Methods* 2015;12:733–735.
26. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* 2014;9:e112963.
27. Li H, Durbin R. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* 2010;26:589–595.
28. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009;25:2078–2079.
29. Vaser R, Sović I, Nagarajan N, Šikić M. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res* 2017;27:737–746.
30. 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.
31. Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 2014;30:2068–2069.
32. Arndt D, Grant JR, Marcu A, Sajed T, Pon A et al. PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic Acids Res* 2016;44:W16–W21.
33. Shaaban S, Cowley LA, McAteer SP, Jenkins C, Dallman TJ et al. Evolution of a zoonotic pathogen: investigating prophage diversity in enterohaemorrhagic *Escherichia coli* O157 by long-read sequencing. *Microb Gen* 2016;2:e000096.
34. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH et al. Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol* 2016;17:132.
35. Rambaut A, Drummond AJ. FigTree; 2018. <https://github.com/rambaut/figtree>
36. Dallman TJ, Ashton P, Schafer U, Ironkin A, Painset A et al. Snap-DB: a database solution for routine sequencing analysis of bacterial isolates. *Bioinformatics* 2018;34:3028–3029.
37. Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA et al. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Res* 2015;43:e15.
38. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 2014;30:1312–1313.
39. Sullivan MJ, Petty NK, Beatson SA. Easyfig: a genome comparison visualizer. *Bioinformatics* 2011;27:1009–1010.
40. Schutz K, Cowley LA, Shaaban S, Carroll A, McNamara E et al. Evolutionary context of non-sorbitol-fermenting Shiga toxin-producing *Escherichia coli* O55:H7. *Emerg Infect Dis* 2017;23:1966–1973.
41. Scotland SM, Willshaw GA, Smith HR, Rowe B. Properties of strains of *Escherichia coli* belonging to serogroup O 157 with special reference to production of Vero cytotoxins VT1 and VT2. *Epidemiol Infect* 1987;99:613–624.
42. 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.
43. Gobin M, Hawker J, Cleary P, Inns T, Gardiner D et al. National outbreak of Shiga toxin-producing *Escherichia coli* O157:H7 linked to mixed salad leaves, United Kingdom, 2016. *Euro Surveill* 2018;23:17-00197.



Analysis Shiga Toxin-Encoding Bacteriophage in a Rare Strain of Shiga Toxin-Producing *Escherichia coli* O157:H7 *stx2a/stx2c*

David R. Greig^{1,2}, Amy F. W. Mikhail¹, Timothy J. Dallman^{1,2} and Claire Jenkins^{1*}

¹ National Infection Service, Public Health England, London, United Kingdom, ² Division of Infection and Immunity, The Royal (Dick) School of Veterinary Studies, The Roslin Institute, The University of Edinburgh, Easter Bush, United Kingdom

OPEN ACCESS

Edited by:

Grzegorz Węgrzyn,
University of Gdańsk, Poland

Reviewed by:

Edward G. Dudley,
Pennsylvania State University (PSU),
United States
James L. Bono,
United States Department of
Agriculture, United States
Angelika Fruth,
Robert Koch Institute (RKI), Germany

*Correspondence:

Claire Jenkins
claire.jenkins1@phe.gov.uk

Specialty section:

This article was submitted to
Virology,
a section of the journal
Frontiers in Microbiology

Received: 29 June 2020

Accepted: 09 September 2020

Published: 21 October 2020

Citation:

Greig DR, Mikhail AFW,
Dallman TJ and Jenkins C (2020)
Analysis Shiga Toxin-Encoding
Bacteriophage in a Rare Strain
of Shiga Toxin-Producing *Escherichia coli* O157:H7 *stx2a/stx2c*.
Front. Microbiol. 11:577658.
doi: 10.3389/fmicb.2020.577658

In December 2015, six cases of Shiga toxin (Stx)-producing *Escherichia coli* (STEC) O157:H7 *stx2a/stx2c* phage type (PT) 24 were identified by the national gastrointestinal disease surveillance system at Public Health England (PHE). Frozen grated coconut imported from India was implicated as the vehicle of infection. Short and long read sequencing data were interrogated for genomic markers to provide evidence that the outbreak strain was from an imported source. The outbreak strain belonged to a sub-lineage (IIa) rare in domestically acquired infection in the United Kingdom, and indicative of an imported strain. Phylogenetic analysis identified the most closely related isolates to the outbreak strain were from cases reporting recent travel not to India, but to Uganda. Phylo-geographical signals based on travel data may be confounded by the failure of local and/or global monitoring systems to capture the full diversity of strains in a given country. This may be due to low prevalence strains circulating in-country under the surveillance radar, or a recent importation event involving the migration of animals and/or people. Comparison of *stx2a*-encoding prophage harbored by the outbreak strain with publicly available *stx2a*-encoding prophage sequences revealed that it was most closely related to *stx2a*-encoding prophage acquired by STEC O157:H7 that caused the first outbreak of STEC-hemolytic uremic syndrome (HUS) in England in 1982–83. Animal and people migration events may facilitate the transfer of *stx2a*-encoding prophage from indigenous STEC O157:H7 to recently imported strains, or vice versa. Monitoring the global transmission of STEC O157:H7 and tracking the exchange of *stx2a*-encoding phage between imported and indigenous strains may provide an early warning of emerging threats to public health.

Keywords: Shiga toxin-producing *E. coli*, outbreak, genomics, public health, epidemiology, coconut

INTRODUCTION

Outbreaks of foodborne, gastrointestinal disease caused by Shiga toxin (Stx)-producing *Escherichia coli* (STEC) serotype O157:H7 are regarded as a significant threat to public health (Riley et al., 1983; Michino et al., 1999; Cowley et al., 2016; Launders et al., 2016a; Gobin et al., 2018). A subset of vulnerable patients at the extremes of age are at risk of developing hemolytic uremic syndrome

(HUS), a condition characterized by renal failure, and cardiac and neurological complications that can be fatal (Tarr et al., 2005; Launders et al., 2016a). STEC O157:H7 is zoonotic and can be transmitted to humans via direct contact with animals and/or their environment, contaminated food or water, or person-to-person spread (Byrne et al., 2015).

There are three main lineages of STEC O157:H7 (I, II, and I/II) and eight sub-lineages (Ia, Ib, Ic, IIa, IIb, IIc, I/IIa, and I/IIb) (Dallman et al., 2015a). With respect to the clade typing scheme proposed by Iyoda et al. (2014), as described previously, lineage I corresponded to clades 1 through 6, lineage II corresponded to clade 7, and lineage I/II corresponded to clade 8 (Eppinger et al., 2011; Dallman et al., 2015a). The majority of human cases and outbreaks of STEC O157:H7 in the United Kingdom are caused by sub-lineages Ic, I/IIa, and IIc (Dallman et al., 2015a). The sub-lineages that are infrequently isolated from human cases in the United Kingdom are more likely to be associated with returning travelers or outbreaks associated with contaminated food imported to the United Kingdom from elsewhere (Gobin et al., 2018).

The defining pathogenicity factor for STEC O157:H7 is the presence of *Stx*-producing genes (*stx*). There are two types of *Stx*, *Stx1* and 2, and at least 10 subtypes (1a–1d and 2a–2i) (Scheut et al., 2012). The genes encoding the *stx* subtypes are located on mobile genetic elements (bacteriophages) that may be acquired by STEC and integrated into the genome. STEC O157:H7 harboring the *stx2a*-encoded phage are most commonly associated with causing HUS (Persson et al., 2007; Byrne et al., 2020).

In December 2015, six cases of STEC O157:H7 phage type (PT) 24 were identified by the national gastrointestinal disease surveillance system at Public Health England (PHE). Following the analysis of the trawling questionnaires, a specific brand of imported frozen grated coconut was implicated as the vehicle of infection. In this study, we conducted an epidemiological investigation of the outbreak, and analyzed short and long read sequencing data to look for genomic markers to provide further evidence that the outbreak strain was from a non-domestic (non-UK) source.

MATERIALS AND METHODS

Microbiological and Epidemiological Data Collection

In England, all fecal specimens from hospital and community cases of gastrointestinal disease submitted to local hospital laboratories are tested for *E. coli* O157:H7. All isolates are submitted to the Gastrointestinal Bacterial Reference Unit (GBRU) at PHE for identification and phage typing. Since July 2015, all isolates have been whole genome sequenced for routine surveillance (National Center for Biotechnology Information Short Read Archive BioProject PRJNA315192).

The majority of isolates of STEC O157:H7 included in this study were submitted to GBRU between July 2015 (when WGS was first implemented) and December 2016 (12 months after the outbreak occurred). During this time frame, 1103

isolates were submitted to GBRU; 149 belonged to household outbreak, 283 were linked to known outbreaks, and 671 were identified as sporadic cases. Additional isolates sequenced from the archive collection submitted to GBRU between 2006 and 2016, included eight isolates belonging to the same PT as the outbreak strain, PT24, all of which reported recent (within 7 days of onset of symptoms) travel to Uganda, and all cases reporting recent travel to the Indian Sub-Continent (ISC; $n = 56$) (Supplementary Table S1).

Epidemiological Data Collection

In January 2009, PHE implemented the National Enhanced Surveillance System for STEC (NESSS) in England. This system has been described in detail previously (Byrne et al., 2015). Briefly, it captures standardized epidemiological on all cases of STEC reported in England through an Enhanced Surveillance Questionnaire (ESQ) including detailed demographic, clinical, and exposure data which is reconciled with microbiological data in NESSS. Following analysis of the ESQ data, each case was re-interview using a standardized trawling questionnaire.

Short Read Sequencing on the Illumina HiSeq 2500

Genomic DNA was extracted from cultures of STEC O157:H7 using the Qiagen Qiasymphony (Qiagen, Hilden, Germany). The sequencing library was prepared using the Nextera XP kit (Illumina, San Diego, CA, United States) for sequencing on the Illumina HiSeq 2500 (Illumina, San Diego, CA, United States) instrument run with the fast protocol. High quality trimmed (leading and trailing trimming at $< Q30$ using Trimmomatic v0.27 (Bolger et al., 2014)). Illumina FASTQ reads were mapped to the Sakai STEC O157 reference genome (NC 002695.1) using BWA MEM v0.7.13 (Li and Durbin, 2010) and Samtools v (Li et al., 2009). Variant positions were identified by GATK v2.6.5 UnifiedGenotyper (McKenna et al., 2010) that passed the following parameters: $> 90\%$ consensus, minimum read depth of 10, Mapping Quality (MQ) ≥ 30 . Any variants called at positions

TABLE 1 | Diversity and frequency of sampling within each sub-lineage 2015–2016.

Sub-lineage	Total number of isolates	Number of sporadic isolates	Number of clusters detected at the 250 SNP level
Ia	31	29	9
Ib	25	25	4
Ic	234	183	4
IIa	325	98	28
IIb	139	44	4
IIc	285	246	1
I/II	37	24	3
NSF	6	4	2
SF	21	18	5
Total	1103	671	

NSF, non-sorbitol fermenting STEC O157:H7; SF, sorbitol fermenting STEC O157:H7.

TABLE 2 | Age and sex ratio and characteristics of isolates from UK travelers returning from the ISC and Uganda 2006–2016 (number in parentheses).

Country	Total	Male	Female	Adult	Child	Lineage (n)	Stx profile (n)
Bangladesh	4	3	1	2	2	Ila (4)	Stx2c (4)
Bhutan	1	0	1	1	0	Ib (1)	Stx2c (1)
India	32	19	16	24	11	Ila (32)	Stx2c (32)
Maldives	0	–	–	–	–		–
Nepal	0	–	–	–	–		–
Pakistan	16	6	9	2	14	Ia (1) Ila (15)	Stx1a/Stx2a Stx2c (15)
Sri Lanka	3	1	2	2	1	Ila (3)	Stx2c (3)
Total ISC	56	29	29	31	28		
Uganda	8	2	6	6	2	Ila (7) Ib (1)	Stx2a/stx2c (4)

TABLE 3 | Size, position, and integration sites of all 19 prophages (P) and six prophage-like elements (PLE) within sample 194195.

Prophage detected by PHASTER	Gene 5' to prophage	Gene 3' to prophage	Size (bp)	Position
P1	<i>cpxF^{Pb}</i>	<i>flieF</i>	31,356	237,188–268,544
P2 ^a	<i>zur</i>	<i>aphA</i>	19,508	433,909–453,417
PLE1	<i>leuX</i>	Hypothetical ^b	9,862	678,235–688,097
PLE2	<i>intA^b</i>	<i>nanM</i>	33,695	688,557–722,252
P3 ^a	tRNA-Thr (cgt)	<i>prgR</i>	54,474	1,128,970–1,183,444
P4	<i>ybhC^b</i>	<i>ybhB</i>	38,693	1,751,265–1,789,958
P5	<i>yccA^b</i>	tRNA-Ser (tga)	49,561	2,021,966–2,071,527
PLE3	<i>ycdU</i>	tRNA-Ser (gga)	85,778	2,170,090–2,255,868
P6 ^a	<i>potC</i>	<i>potB</i>	48,949	2,342,248–2,391,197
P7	<i>potA</i>	<i>caeB^b</i>	65,659	2,393,794–2,459,453
P8 ^a	<i>ompW</i>	<i>ispE</i>	79,722	2,544,993–2,624,715
P9 ^a	<i>ompW</i>	<i>trpA</i>	98,899	2,675,966–2,774,865
P10 ^a	<i>ttcA</i>	<i>rspR</i>	101,788	2,877,961–2,979,749
P11	<i>ompN</i>	<i>rspA^b</i>	45,568	3,160,400–3,205,968
P12	<i>yecA^b</i>	tRNA-Leu (taa)	20,043	3,542,644–3,562,687
P13	<i>yodB</i>	tRNA-Ser (cga) ^b	43,687	3,617,610–3,661,297
PLE4	<i>cobU</i>	Hypothetical	13,348	3,684,489–3,697,837
P14 (stx2c)	<i>yeeW</i>	<i>sbvB^b</i>	60,586	3,698,735–3,759,321
P15 ^a	<i>yehV</i>	<i>mlrA^b</i>	52,982	3,900,672–3,953,654
P16 (stx2a)	<i>yfcV</i>	<i>argW^b</i>	74,662	4,190,350–4,265,012
P17	<i>argW^b</i>	<i>lacY</i>	8,233	4,265,444–4,273,677
P18	<i>ssrA</i>	<i>alpA^b</i>	20,904	4,549,827–4,570,731
PLE5	tRNA-Phe (gaa)	<i>pitB</i>	23,331	4,924,934–4,948,265
P19	<i>yicC^b</i>	<i>yicC</i>	15,440	5,631,805–5,647,245
PLE6 (LEE)	<i>selC</i>	<i>yicL</i>	43,021	5,665,860–5,708,881

^ARefers to prophages that appear to be compound prophages (i.e., two or more prophages that are sequential or one phage integrated inside another, with intact integrase genes). ^BRefers to the end in which the Integrase gene (*intA*) is located.

that were within the known prophages in Sakai were masked from further analyses. The remaining variants were imported into SnapperDB v0.2.5 (Dallman et al., 2018).

Hierarchical single linkage clustering was performed on the pairwise single nucleotide polymorphisms (SNPs) difference between all strains at various distance thresholds (250, 100, 50, 25, 10, 5, 0). The result of the clustering is an SNP profile, or SNP address, that can be used to describe the population structure based on clonal groups (Dallman et al., 2015b, 2018). Although isolates greater than 5 SNPs apart are unlikely to be part of the same temporally linked outbreak, deeper phylogenetic relationships within the 10 or 25 SNP clusters may provide epidemiologically useful information or associations. Lineage and

sub-lineage assignment were performed based on discriminatory SNPs, extracted directly from SnapperDB v0.2.5, that define the population structure, as described previously (Dallman et al., 2015b, 2018).

Long Read Sequencing Using ONT and Data Processing

Genomic DNA was extracted and purified using the Qiagen Genomic Tip, midi 100/G (Qiagen, Hilden, Germany) with minor alterations including no vigorous mixing steps (performed by inversion) and elution into 100 µl. DNA was quantified using a Qubit and the high sensitivity (HS) dsDNA Assay

TABLE 4 | Summary of the PHE archived and publicly available strains used within this study, their Strain ID, lineage, phage type, *stx* profile, assembly accession numbers, and NCBI BioProject.

Strain ID	lineage	Phage type	<i>Stx</i> profile	Number of prophages	Reference	BioProject #	Assembly accession #
PHE archive							
194195	IIa	PT24	<i>Stx2a/2c</i>	19	This study	PRJNA315192	CP044350
E30228	Ia	PT4	<i>Stx1a/2a</i>	15	Scotland et al., 1987	PRJNA315192	VXJO00000000
E34500	I/IIa	PT2	<i>Stx2a/2c</i>	14	Taylor et al., 1986	PRJNA315192	VXJN00000000
E45000	I/IIb	PT49	<i>Stx2a</i>	17	Yara et al., 2020	PRJNA315192	VXJM00000000
E116508	Ic	PT21/28	<i>Stx2a/2c</i>	17	Yara et al., 2020	PRJNA315192	VXJP00000000
315176	IIb	PT8	<i>Stx2a</i>	16	Byrne et al., 2018	PRJNA315192	VXJQ00000000
267849	IIa	PT34	<i>Stx2a/2c</i>	16	Gobin et al., 2018	PRJNA315192	VXJR00000000
Publicly available							
9000	Ic	PT21/28	<i>Stx2a/2c</i>	17	Shaaban et al., 2016	PRJNA336330	CP018252
397404	Ic	PT21/28	<i>Stx2a/2c</i>	15	Yara et al., 2020	PRJNA315192	CP043019
155	Ic	PT32	<i>Stx2a</i>	18	Shaaban et al., 2016	PRJNA336330	CP018237
350	IIc	PT8	<i>Stx1a/2c</i>	16	Launders et al., 2016b	PRJNA336330	CP018243
272	I/IIa	PT2	<i>Stx2a</i>	16	Jenkins et al., 2015	PRJNA336330	CP018239
644	IIc	PT8	<i>Stx1a/2c</i>	18	Cowley et al., 2016	PRJNA321984	CP015831
180	IIc	PT54	<i>Stx1a/1a/2c</i>	15	Cowley et al., 2016	PRJNA321984	CP015832
Sakai	Ia	n/a	<i>Stx1a/1a/2a</i>	18	Michino et al., 1999	PRJNA57781	NC_002695
EDL933	Ia	n/a	<i>Stx1a/2a</i>	14	Riley et al., 1983	PRJNA253471	CP008957
EC4115	I/IIa	n/a	<i>Stx2a/2c</i>	17	Uhlich et al., 2008	PRJNA224116	NC_011353
TW14359	I/IIa	n/a	<i>Stx2a/2c</i>	17	Uhlich et al., 2008	PRJNA224116	NC_013008

Kit (ThermoFisher Scientific, Waltham, MA, United States) to manufacturer's instructions. Library preparation was performed using the Native Barcoding kit (SQK-LSK108 and EXP-NBD103) (Oxford Nanopore Technologies, Oxford, United Kingdom). The prepared library was loaded on a FLO-MIN106 R9.4.1 flow cell (Oxford Nanopore Technologies, Oxford, United Kingdom) and sequenced using the MinION for 24 h.

Data produced in a raw FAST5 format were basecalled into FASTQ format and de-multiplexed using Guppy v2.3.5 (Oxford Nanopore Technologies). The reads were then de-multiplexed again using Deepbiner v0.2.0 (Wick et al., 2018) to reduce barcode contamination. Run metrics were generated using Nanoplot v1.8.1 (De Coster et al., 2018). The barcode and y-adaptor from each sample's reads were trimmed, and chimeric reads split using Porechop v0.2.4. Finally, trimmed reads were filtered using Filtlong v0.1.1 with the following parameters; min_length = 1000, keep_percent = 90, and target_bases = 275Mb, to generate approximately 50x coverage of the STEC genome with the longest and highest quality reads.

De novo Assembly, Polishing, Reorientation, and Annotation

Trimmed ONT FASTQ files were assembled using Canu v1.7.1 (Koren et al., 2017). Polishing of the assembly was performed in a three-step process first, using Nanopolish v0.11.1 (Loman et al., 2015) using both the trimmed ONT FASTQs and FAST5s for each respective sample accounting for methylation using the -methylation-aware = dam,dcm and -min-candidate-frequency = 0.1. Secondly, Pilon v1.22 (Walker et al., 2014) using Illumina FASTQ reads as the query dataset with the use

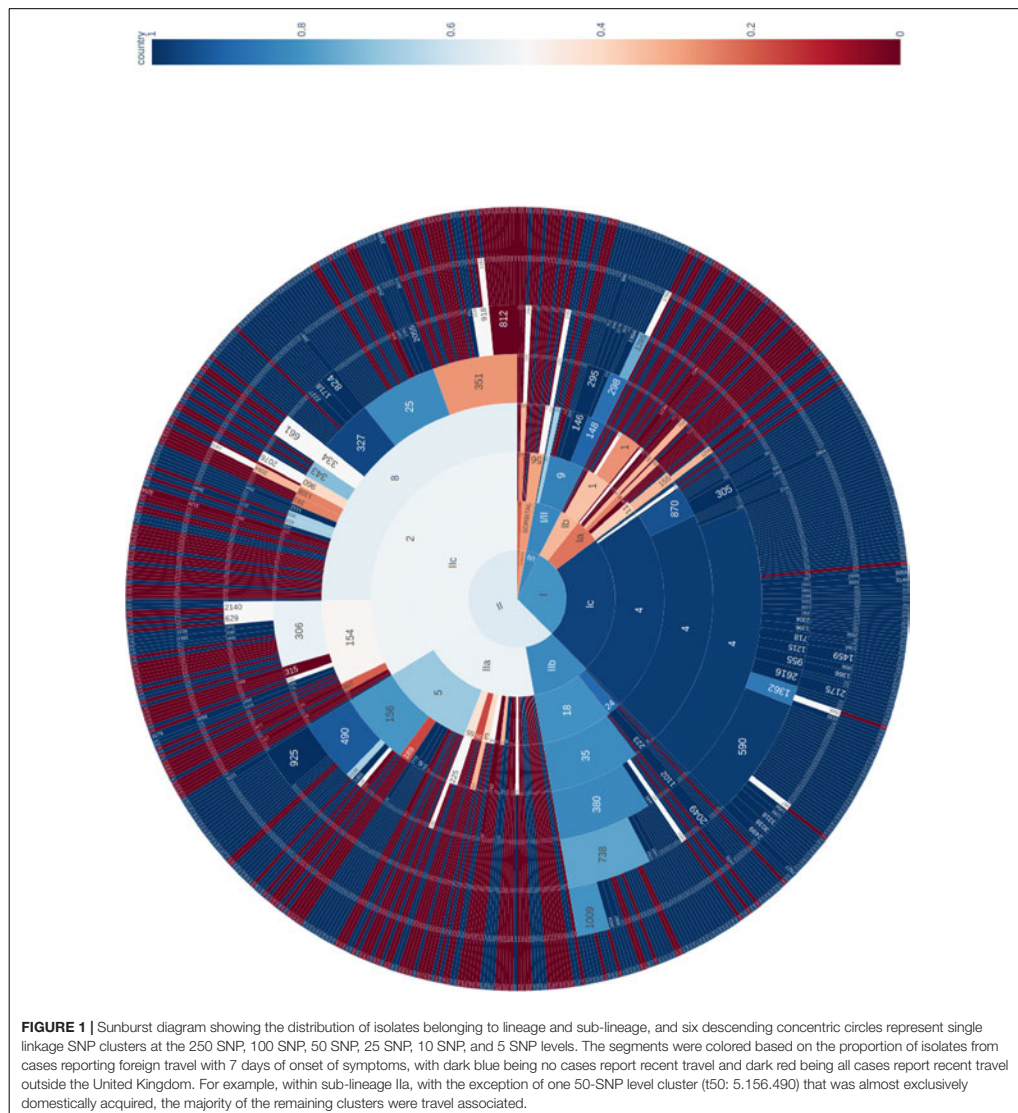
of BWA v0.7.17 (Li and Durbin, 2010) and Samtools v1.7 (Li et al., 2009). Finally, Racon v1.2.1 (Vaser et al., 2017) also using BWA v0.7.17 and Samtools v1.7 (Li et al., 2009) was used with the Illumina reads to produce a final assembly for each of the samples. As the chromosome from the assembly was closed, it was re-orientated to start at the *dnaA* gene (NC_000913) from *E. coli* K-12, using the -fixstart parameter in circlator v1.5.5 (Hunt et al., 2015). Prokka v1.13 was used to annotate the draft assembly (Seemann, 2014).

Prophage Detection, Excision, and Processing

Prophages across both samples were detected and extracted using the Phage Search Tool (PHASTER) (Arndt et al., 2016). Prophage extraction from the genome occurred regardless of prophage size or quality and any detected prophages separated by less than 4 kbp were conjoined into a single phage using Propi v0.9.0 as described in Shaaban et al. (2016). From here the prophages were manually trimmed to remove any non-prophage genes and were again annotated using Prokka v 1.13 with the use of a personalized database (Amino Acid multi-FASTA) containing known STEC prophage genes was used to annotate the final assemblies. Database publicly available from https://github.com/gingerdave269/prophage_DB. The output GenBank (gbk) files were modified to color genes by function.

Mash and Phylogeny

Mash v2.2 (Ondov et al., 2016) was used to sketch (sketch length 1000, kmer length, 21) *stx*-encoding prophages from the outbreak isolate sequenced using ONT in this study, and the publicly available STEC genomes listed in Table 4. The pairwise Jaccard



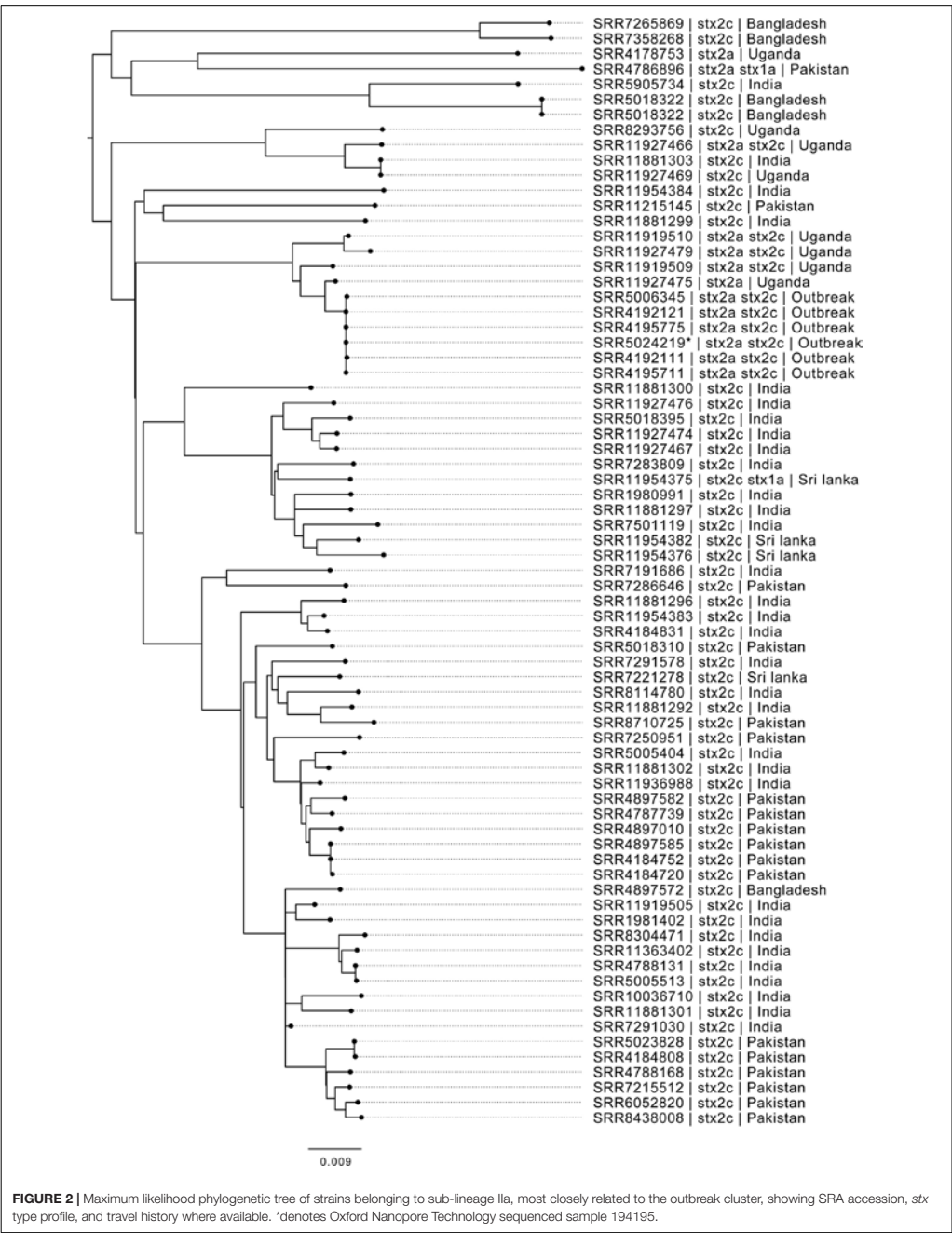
distance between the prophages was calculated and a neighbor joining tree computed and visualized using FigTree v1.4.4.

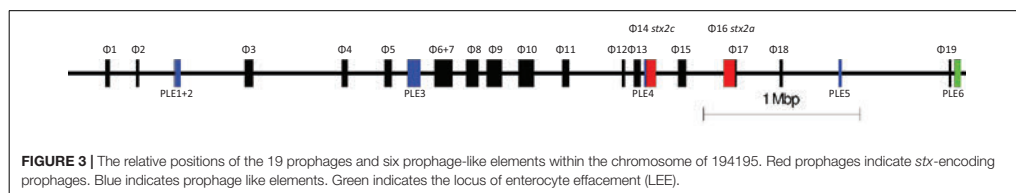
Visualization Tools

All gene diagrams were constructed using Easyfig v2.2.3 (Sullivan et al., 2011). Neighbor joining trees were visualized and annotated using FigTree v1.4.4.

Data Deposition

Illumina FASTQ files for all samples used in the study can be found under BioProject PRJNA315192. Nanopore FASTQ file is available under SRA accession: SRR10177137. The assembly can be found under the following accession: CP044350. All the above are available from BioProject: PRJNA315192.





RESULTS AND DISCUSSION

Epidemiological Investigations

In December 2015, the Gastrointestinal Bacteria Reference Unit identified six cases of STEC O157:H7 with a rare PT, PT24. Whole genome sequencing results confirmed that the cases belong to the same 5-SNP single linkage cluster. The outbreak strain belonged to sub-lineage IIa and had the *stx* profile *stx2a* and *stx2c*.

An outbreak control team was convened, and trawling questionnaires were completed for each case and their symptomatic household contacts. Of the six cases, two were male and four were female, and the age range for all six cases was 1–45 years old. Dates of onset of symptoms of gastrointestinal infection fell between 17 and 25 November 2015. The cases were national distributed. Analysis of the trawling questionnaires identified consumption of frozen grated coconut belonging to the same brand was reported by four of the six cases. Of the remaining two cases, one patient recalled eating coconut yogurt, while the other reported the consumption of Indian sweets and attended a Diwali party in the days before onset of symptoms but did not recall specifically eating coconut.

Microbiological analysis of an unopened packet of the coconut product from the restaurant where two of the cases ate, detected unsatisfactory levels of *E. coli* ($>10^2$ cfu/g), as defined by Regulation (EC) No. 178/2002¹, although no STEC O157:H7 was detected in the sample.

Phylogenetic Analysis of WGS Data Linked to Patients Travel History to Determine a Domestic or Non-domestic Origin for the Outbreak Strain

Following the epidemiological investigation that provided evidence that the contaminated food vehicle may be imported, we analyzed the genome data to look for evidence that the outbreak strain had a non-domestic origin.

Of the 671 sporadic isolates of STEC O157:H7 referred to GBRU between July 2015 and December 2016, the majority belonged to sub-lineages IIc (246/671, 36.7%) and Ic (183/671, 27.2%) (Table 1). In contrast, sub-lineage IIa was less frequently isolated from sporadic cases (98/671, 14.6%), and showed the highest level of sub-lineage diversity, as measured by the number

of different 250-SNP single linkage clusters within each sub-lineage (Table 1). High levels of sub-lineage diversity are representative of infrequent sampling from a geographically dispersed reservoir (Gobin et al., 2018).

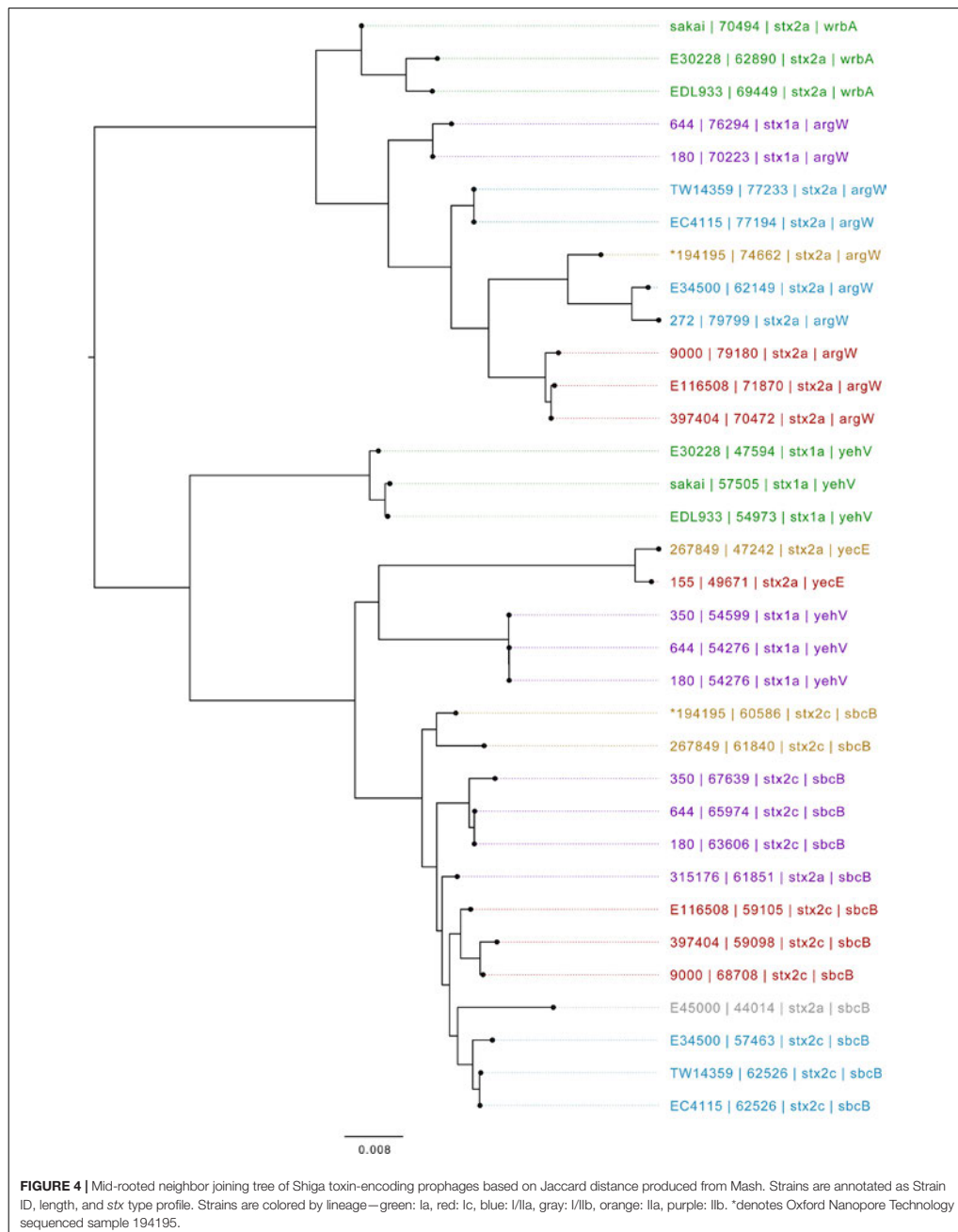
The phylogeny of the 671 sporadic isolates included in this study were visualized using a sunburst diagram showing the distribution of isolates belonging to lineage and sub-lineage, and six descending single linkage SNP clusters at the 250 SNPs, 100 SNPs, 50 SNPs, 25 SNPs, 10 SNPs, and 5 SNPs level (Figure 1) (Dallman et al., 2018). In Figure 1, the inner circle shows the proportions of the three main lineages and the proportion of those isolates that fall outside the lineage structure. Moving outward from the inner circle, the second concentric circle shows the proportion of isolates in each sub-lineage, and the third, fourth, fifth, sixth, and seventh concentric circles from the inner circle represent the 250 SNPs, 100 SNPs, 50 SNPs, 25 SNPs, 10 SNPs, and 5 SNPs levels, respectively. The numbers represent SNP type or SNP address designation, for example, the SNP address designation for the largest 25 SNP single linkage cluster in sub-lineage IIa is t25: 5.156.490.925%.

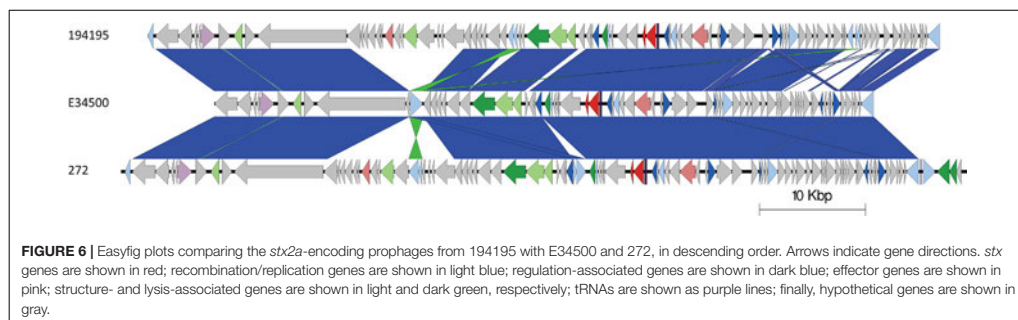
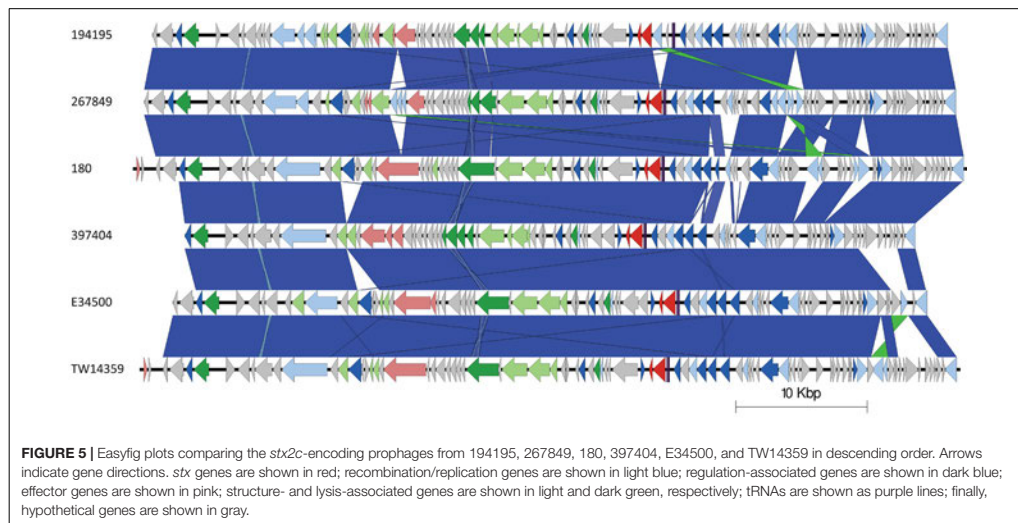
The segments were colored based on the proportion of isolates from cases reporting foreign travel with 7 days of onset of symptoms, with dark blue being no cases report recent travel and dark red being all cases report recent travel outside the United Kingdom. The majority of clusters that belong to sub-lineages Ic, IIb, and I/II were domestically acquired. These data were consistent with previous studies that indicated these sub-lineages were likely to be endemic in the United Kingdom (Dallman et al., 2015a; Adams et al., 2016; Byrne et al., 2018). Sub-lineages IIa and IIc displayed a mixture of domestically acquired and travel associated clusters (Figure 1). However, within each sub-lineage, at more discriminatory SNP levels, clear cluster associations to either travel or domestic-acquisition were observed. For example, within sub-lineage IIa, with the exception of one 50-SNP level cluster (t50: 5.156.490.) that was almost exclusively domestically acquired, the majority of the remaining clusters were travel associated. The outbreak strain described in this study fell within the travel-associated clusters providing evidence that the contaminated food vehicle was most likely from an imported source.

Analysis of WGS Data of Isolates From UK Residents Reporting Recent Travel to the Indian Sub-Continent

A closer look at the phylogenetic context of the outbreak strain revealed the most closely related strains were isolates from

¹<https://www.gov.uk/government/publications/ready-to-eat-foods-microbiological-safety-assessment-guidelines>





cases reporting recent travel to Uganda (Figure 2). As the epidemiological investigation indicated that the country of origin of the outbreak strains was India, we included sequences of all the isolates of STEC O157:H7 in the PHE database isolated from travelers recently returned to the United Kingdom from the ISC (Figure 2).

Interrogation of the PHE database of all isolates submitted to GBRU between 2006 and 2016 ($n = 11,339$), identified 56 isolates from UK travelers recently returned from the ISC, the majority reporting recent travel to India ($n = 32$) or Pakistan ($n = 16$) (Table 2). The majority of isolates from the ISC belonged to sub-lineage IIa (53/56), and all had *stx2c* only ($n = 56$). Evidence from studies testing samples from humans, food, and animals in India indicates that the prevalence of STEC O157:H7 is low in this country (Sehgal et al., 2008; Lanjewar et al., 2010; Shrivastava et al., 2017).

There were eight isolates from UK travelers recently returned from Uganda, all but one belonged to sub-lineage IIa

(7/8). Of these, 4/8 had the same *stx* profile (*stx2a/stx2c*) as the outbreak strain, and all belonged to a unique clade that had acquired a *stx2a*-encoding bacteriophage. Evidence for the prevalence of STEC O157:H7 in East Africa from human clinical specimens is sparse although STEC O157:H7 has been detected in a number of studies sampling the animal reservoir in this region (Kaddu-Mulindw et al., 2001; Grace et al., 2008; Dulo et al., 2015; Beyi et al., 2017).

Phylo-geographical signals have proved useful in providing evidence for the likely country of origin of STEC and *Salmonella* causing outbreaks of foodborne gastrointestinal disease (Allard et al., 2016; Cowley et al., 2016; Hoffmann et al., 2016; Gobin et al., 2018; Pijnacker et al., 2019). However, there are factors that may confound this signal. Strains of STEC O157:H7 isolated from returning travelers only provide a snapshot of strains endemic to the country visited and may not reflect the full diversity of strains circulating in a given region. Although the outbreak strain may be

associated with travel to Uganda, it may also be circulating at low prevalence in India, perhaps due to a recent importation event.

Analysis of the Outbreak Strain Using Oxford Nanopore Technology

We analyzed long read sequencing data of the outbreak strain, primarily to determine the sequence and genomic structure of the *stx2*-encoding bacteriophage. The outbreak isolate assembled into a chromosome of 5,753,376 bp and a single plasmid, pO157 (111,625 bp, IncFIB). The genome of the outbreak isolate comprised 19 prophages (Table 3 and Figure 3).

Based on Mash distance, the *stx2c*-encoding prophage located at the *stx*-encoding bacteriophage insertion (SBI) site *sbcB*, clustered on the same branch as *stx2c* prophage from strains from different time frames, geographical regions, and sub-lineages, most closely related to other *stx2c*-encoding prophage from sub-lineage IIa (Figure 4). The *stx2c*-encoding prophage from the outbreak strains aligned across the length of other *stx2c*-encoding prophage with few structural variations (Figure 5). The conserved nature of the *stx2c*-encoding bacteriophage was consistent with the hypothesis that the *stx2c*-encoding bacteriophage was acquired prior to the global dissemination and regional expansions of STEC O157:H7 (Dallman et al., 2015a).

The *stx2a*-encoding prophage acquired by the outbreak strain was located at SBI site *argW*. Compared to the *stx2c*-encoding prophage, the *stx2a*-encoding prophage exhibits greater diversity, based on Mash distance and whole prophage alignment (Figure 4). In certain countries, the regional expansion of specific sub-lineages, or clades within sub-lineages, has involved acquisition of a *stx2a*-encoding prophage (Dallman et al., 2015a; Yara et al., 2020).

The *stx2a*-encoding prophage from the outbreak strain was most closely related to *stx2a*-encoding prophage acquired by a strain belonging to lineage I/II that caused the first outbreak of STEC-HUS in the West Midlands in England in the early 1980s (Figures 4, 6) (Taylor et al., 1986; Yara et al., 2020). Previous studies have shown that this lineage I/II strain of STEC O157:H7 harboring *stx2c*-encoding prophage had been indigenous in UK domestic cattle population for decades, and that it emerged as a threat to public health in 1983, following the acquisition of a *stx2a*-encoding prophage at some point during the previous decade (Dallman et al., 2015a; Yara et al., 2020).

CONCLUSION

The use of WGS for surveillance of gastrointestinal pathogens enabled us to identify a small, geographically dispersed outbreak of foodborne disease. The epidemiological analysis provided evidence that the outbreak strains originated from India, while the phylogenetic analysis of the sequencing data indicated the strain was most closely related to isolates from Uganda, and the *stx2a*-encoding phage was most closely related to *stx2a*-encoding bacteriophage harbored by the strains of STEC O157:H7 that emerged in the United Kingdom, as the most common cause of STEC-HUS in early 1980s.

These analyses described in this study are open to interpretation in a number of different ways. Microbiological investigation of the grated coconut samples did not detect STEC O157:H7, and the contaminated food vehicle may have been an imported product from elsewhere. Alternatively, the epidemiological evidence indicating India as the country of origin of the outbreak strain may have been correct, but the phylo-geographical signal was obscured by the low prevalence of the outbreak strain in that region. Strains of STEC O157:H7 that have a low prevalence in a specific region may not be captured by either local or global monitoring systems, and it is likely the full diversity of strains in a given region may circulate under the surveillance radar. Moreover, non-indigenous strains of STEC O157:H7 may be introduced into a new region by the migration of animals (including migratory birds) and people, thus further confounding the phylo-geographical signal. These animal and people migration events may facilitate the transfer of mobile genetic elements, such as the *stx2a*-encoding prophage, from indigenous strains of STEC O157:H7 to recently imported strains, or vice versa. Monitoring the transmission of strains of STEC O157:H7 on a global scale, and tracking the exchange of *stx2a*-encoding phage between imported and indigenous strains, may provide an early warning of emerging threats to public health.

DATA AVAILABILITY STATEMENT

Illumina FASTQ files for all samples used in the study can be found under BioProject PRJNA315192. Nanopore FASTQ file is available under SRA accession: SRR10177137. The assembly can be found under the following accession: CP044350. All the above are available from BioProject: PRJNA315192.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent from the participants' legal guardian/next of kin was not required to participate in this study in accordance with the national legislation and the institutional requirements. The authors declare that there is no requirement for ethical approval for this submission. This work was undertaken to inform the delivery of patient care and to prevent the spread of infection, defined as USUAL PRACTICE in public health and health protection.

AUTHOR CONTRIBUTIONS

AM performed epidemiological investigations. DG performed DNA extraction, library preparation, Nanopore sequencing, data processing, genome assembly, correction, and annotation, created Easyfig diagrams, and performed the prophage comparison using Mash with associated scripts designed

by TD. TD created sunburst plots. CJ and DG wrote the original manuscript. CJ, DG, and TD reviewed the manuscript. CJ and TD supervised DG. All authors contributed to the article and approved the submitted version.

FUNDING

This research was part funded by the National Institute for Health Research (NIHR) Health Protection Research Unit in Gastrointestinal Infections at the University of Liverpool (United Kingdom), in partnership with PHE, in collaboration with the University of East Anglia (United Kingdom), the

University of Oxford (United Kingdom), and the Quadram Institute (United Kingdom). CJ, TD, and DG are based at PHE. The views expressed are those of the authors and not necessarily those of the National Health Service, the NIHR, the Department of Health nor PHE.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fmicb.2020.577658/full#supplementary-material>

REFERENCES

- Adams, N. L., Byrne, L., Smith, G. A., Elson, R., Harris, J. P., Salmon, R., et al. (2016). Shiga toxin-producing *Escherichia coli* O157, England and Wales, 1983–2012. *Emerg. Infect. Dis.* 22, 590–597. doi: 10.3201/eid2204.151485
- Allard, M. W., Strain, E., Melka, D., Bunning, K., Musser, S. M., Brown, E. W., et al. (2016). Practical value of food pathogen traceability through building a whole-genome sequencing network and database. *J. Clin. Microbiol.* 54, 1975–1983. doi: 10.1128/JCM.00081-16
- Arndt, D., Grant, J. R., Marcu, A., Sajed, T., Pon, A., Liang, Y., et al. (2016). PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic Acid Res.* 44, W16–W21. doi: 10.1093/nar/gkw387
- Beyi, A. F., Fite, A. T., Tora, E., Tafese, A., Genu, T., Kaba, T., et al. (2017). Prevalence and antimicrobial susceptibility of *Escherichia coli* O157 in beef at butcher shops and restaurants in central Ethiopia. *BMC Microbiol.* 17:49. doi: 10.1186/s12866-017-0964-z
- Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics* 30, 2114–2120. doi: 10.1093/bioinformatics/btu170
- Byrne, L., Adams, N., and Jenkins, C. (2020). Association between Shiga toxin-producing *Escherichia coli* O157:H7 *stx* gene subtype and disease severity, England, 2009–2019. *Emerg. Infect. Dis.* 26, 2394–2400. doi: 10.3201/eid2610.200319
- Byrne, L., Dallman, T. J., Adams, N., Mikhail, A. F. W., McCarthy, N., and Jenkins, C. (2018). Highly pathogenic clone of shiga toxin-producing *Escherichia coli* O157:H7, England and Wales. *Emerg. Infect. Dis.* 24, 2303–2308. doi: 10.3201/eid2412.180409
- Byrne, L., Jenkins, C., Launders, N., Elson, R., and Adak, G. K. (2015). The epidemiology, microbiology and clinical impact of Shiga toxin-producing *Escherichia coli* in England, 2009–2012. *Epidemiol. Infect.* 143, 3475–3487. doi: 10.1017/S0950268815000746
- Cowley, L. A., Dallman, T. J., Fitzgerald, S., Irvine, N., Rooney, P. J., McAteer, S. P., et al. (2016). Short-term evolution of Shiga toxin-producing *Escherichia coli* O157:H7 between two food-borne outbreaks. *Microb. Genom.* 2:e000084. doi: 10.1099/mgen.0.000084
- Dallman, T. J., Ashton, P., Schafer, U., Jironkin, A., Painset, A., Shaaban, S., et al. (2018). SnapperDB: a database solution for routine sequencing analysis of bacterial isolates. *Bioinformatics* 34, 3028–3029. doi: 10.1093/bioinformatics/bty212
- Dallman, T. J., Ashton, P. M., Byrne, L., Perry, N. T., Petrovska, L., and Ellis, R. (2015a). Applying phylogenomics to understand the emergence of Shiga-toxin-producing *Escherichia coli* O157:H7 strains causing severe human disease in the UK. *Microb. Genom.* 1:e000029. doi: 10.1099/mgen.0.000029
- Dallman, T. J., Byrne, L., Ashton, P. M., Cowley, L. A., Perry, N. T., and Adak, G. (2015b). Whole-genome sequencing for national surveillance of Shiga toxin-producing *Escherichia coli* O157. *Clin. Infect. Dis.* 61, 305–312. doi: 10.1093/cid/civ318
- De Coster, W., D'Hert, S., Scheut, D. T., Cruts, M., and Van Broeckhoven, C. (2018). NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics* 34, 2666–2669. doi: 10.1093/bioinformatics/bty149
- Dulo, F., Feleke, A., Szonyi, B., Fries, R., Baumann, M. P., and Grace, D. (2015). Isolation of multidrug-resistant *Escherichia coli* O157 from goats in the somali region of ethiopia: a cross-sectional, abattoir-based study. *PLoS One* 10:e0142905. doi: 10.1371/journal.pone.0142905
- Eppinger, M., Mammel, M. K., Leclerc, J. E., Ravel, J., and Cebula, T. A. (2011). Genomic anatomy of *Escherichia coli* O157:H7 outbreaks. *Proc Natl. Acad. Sci. U S A* 108, 20142–20147. doi: 10.1073/pnas.1107176108
- Gobin, M., Hawker, J., Cleary, P., Inns, T., Gardiner, D., and Mikhail, A. (2018). National outbreak of Shiga toxin-producing *Escherichia coli* O157:H7 linked to mixed salad leaves, United Kingdom, 2016. *Euro Surveill.* 23:e0017–197. doi: 10.2807/1560-7917.ES.2018.23.18.17-00197
- Grace, D., Omore, A., Randolph, T., Kang'ethe, E., Nasinyama, G. W., and Mohammed, H. O. (2008). Risk assessment for *Escherichia coli* O157:H7 in marketed unpasteurized milk in selected East African countries. *J. Food Prot.* 71, 257–263. doi: 10.4315/0362-028x-71.2.257
- Hoffmann, M., Luo, Y., Monday, S. R., Gonzalez-Escalona, N., Ottesen, A. R., and Muruvanda, T. (2016). Tracing origins of the *Salmonella* bareilly strain causing a food-borne outbreak in the United States. *J. Infect. Dis.* 213, 502–508. doi: 10.1093/infdis/jiv297
- Hunt, M., Silva, N. D., Otto, T. D., Parkhill, J., Keane, J. A., and Harris, S. R. (2015). Circlator: automated circularization of genome assemblies using long sequencing reads. *Genome Biol.* 16:294. doi: 10.1186/s13059-015-0849-0
- Iyoda, S., Manning, S. D., Seto, K., Kimata, K., Isobe, J., Etoh, Y., et al. (2014). Phylogenetic clades 6 and 8 of enterohemorrhagic *Escherichia coli* O157:H7 with particular *stx* subtypes are more frequently found in isolates from hemolytic uremic syndrome patients than from asymptomatic carriers. *Open Forum Infect. Dis.* 1:ofu061. doi: 10.1093/ofid/ofu061
- Jenkins, C., Dallman, T. J., Launders, N., Willis, C., Byrne, L., Jorgensen, F., et al. (2015). Public health investigation of two outbreaks of Shiga toxin-producing *Escherichia coli* O157 associated with consumption of watercress. *Appl. Environ. Microbiol.* 81, 3946–3952. doi: 10.1128/AEM.04188-14
- Kaddu-Mulindw, D. H., Aisu, T., Gleier, K., Zimmermann, S., and Beutin, L. (2001). Occurrence of Shiga toxin-producing *Escherichia coli* in fecal samples from children with diarrhea and from healthy zebu cattle in Uganda. *Int. J. Food Microbiol.* 66, 95–101. doi: 10.1016/s0168-1605(00)00493-1
- Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., and Phillippy, A. M. (2017). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* 27, 722–736. doi: 10.1101/gr.215087.116
- Lanjewar, M., De, A. S., Mathru, M., and Diarrheagenic, E. (2010). Coli in hospitalized patients: special reference to shiga-like toxin producing *Escherichia coli*. *Indian J. Pathol. Microbiol.* 53, 75–78. doi: 10.4103/0377-4929.59188
- Launders, N., Byrne, L., Jenkins, C., Harker, K., Charlett, A., and Adak, G. K. (2016a). Disease severity of Shiga toxin-producing *E. coli* O157 and factors influencing the development of typical haemolytic uraemic syndrome: a retrospective cohort study, 2009–2012. *BMJ Open* 6:e009933. doi: 10.1136/bmjopen-2015-009933
- Launders, N., Locking, M. E., Hanson, M., Willshaw, G., Charlett, A., Salmon, R., et al. (2016b). A large Great Britain-wide outbreak of STEC O157 phage type 8 linked to handling of raw leeks and potatoes. *Epidemiol. Infect.* 144, 171–181. doi: 10.1017/S0950268815001016

- Li, H., and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* 26, 589–595. doi: 10.1093/bioinformatics/btp698
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). 1000 genome project data processing subgroup. The Sequence alignment/map (SAM) format and SAMtools. *Bioinformatics* 25, 2078–2079. doi: 10.1093/bioinformatics/btp352
- Loman, N. J., Quick, J., and Simpson, J. T. (2015). A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nat. Methods* 12, 733–735. doi: 10.1038/nmeth.3444
- McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernysky, A., et al. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20, 1297–1303. doi: 10.1101/gr.107524.110
- Michino, H., Araki, K., Minami, S., Takaya, S., Sakai, N., Miyazaki, M., et al. (1999). Massive outbreak of *Escherichia coli* O157:H7 infection in schoolchildren in Sakai City, Japan, associated with consumption of white radish sprouts. *Am. J. Epidemiol.* 150, 787–796. doi: 10.1093/oxfordjournals.aje.a010082
- Ondov, B. D., Treangen, T. J., Melsted, P., Mallonee, A. B., Bergman, N. H., Koren, S., et al. (2016). Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.* 17:132. doi: 10.1186/s13059-016-0997-x
- Persson, S., Olsen, K. E. P., Ethelberg, S., and Scheut, F. (2007). Subtyping method for *Escherichia coli* shiga toxin (verocytotoxin) 2 variants and correlations to clinical manifestations. *J. Clin. Microbiol.* 45, 2020–2024. doi: 10.1128/jcm.02591-06
- Pijnacker, R., Dallman, T. J., Tijms, A. S. L., Hawkins, G., Larkin, L., and Kotila, S. M. (2019). International Outbreak Investigation Team. An international outbreak of *Salmonella enterica* serotype Enteritidis linked to eggs from Poland: a microbiological and epidemiological study. *Lancet Infect. Dis.* 19, 778–786. doi: 10.1016/S1473-3099(19)30047-7
- Riley, L. W., Remis, R. S., Helgeson, S. D., McGee, H. B., Wells, J. G., Davis, B. R., et al. (1983). Hemorrhagic colitis associated with a rare *Escherichia coli* serotype. *N. Engl. J. Med.* 308, 681–685. doi: 10.1056/NEJM198303243081203
- Scheut, F., Teel, L. D., Beutin, L., Piérard, D., Buvens, G., and Karch, H. (2012). Multicenter evaluation of a sequence-based protocol for subtyping Shiga toxins and standardizing Stx nomenclature. *J. Clin. Microbiol.* 50, 2951–2963. doi: 10.1128/jcm.00860-12
- Scotland, S. M., Willshaw, G. A., Smith, H. R., and Rowe, B. (1987). Properties of strains of *Escherichia coli* belonging to serogroup O 157 with special reference to production of Vero cytotoxins VT1 and VT2. *Epidemiol. Infect.* 99, 613–624. doi: 10.1017/s0950268800066462
- Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 30, 2068–2069. doi: 10.1093/bioinformatics/btu153
- Sehgal, R., Kumar, Y., and Kumar, S. (2008). Prevalence and geographical distribution of *Escherichia coli* O157 in India: a 10-year survey. *Trans. R. Soc. Trop. Med. Hyg.* 102, 380–383. doi: 10.1016/j.trstmh.2008.01.015
- Shaaban, S., Cowley, L., McAteer, S. P., Jenkins, C., Dallman, T. J., Bono, J. L., et al. (2016). Evolution of a zoonotic pathogen: investigating prophage diversity in enterohaemorrhagic *Escherichia coli* O157 by long-read sequencing. *Microb. Gen.* 2:e000096. doi: 10.1099/mgen.0.000096
- Shrivastava, A. K., Kumar, S., Mohakud, N. K., Suar, M., and Sahu, P. S. (2017). Multiple etiologies of infectious diarrhea and concurrent infections in a pediatric outpatient-based screening study in Odisha, India. *Gut. Pathog.* 9:16. doi: 10.1186/s13099-017-0166-0
- Sullivan, M. J., Petty, N. K., and Beatson, S. A. (2011). Easyfig: a genome comparison visualizer. *Bioinformatics* 27, 1009–1010. doi: 10.1093/bioinformatics/btr039
- Tarr, P. I., Gordon, C. A., and Chandler, W. L. (2005). Shiga-toxin-producing *Escherichia coli* and haemolytic uraemic syndrome. *Lancet* 365, 1073–1086. doi: 10.1016/s0140-6736(05)71144-2
- Taylor, C. M., White, R. H., Winterborn, M. H., and Rowe, B. (1986). Haemolytic-uraemic syndrome: clinical experience of an outbreak in the West Midlands. *BMJ* 292, 1513–1516. doi: 10.1136/bmj.292.6534.1513
- Uhlich, G. A., Sinclair, J. R., Warren, N. G., Chmielecki, W. A., and Fraticamo, P. (2008). Characterization of Shiga toxin-producing *Escherichia coli* isolates associated with two multistate food-borne outbreaks that occurred in 2006. *Appl. Environ. Microbiol.* 74, 1268–1272. doi: 10.1128/AEM.01618-07
- Vaser, R., Soviæ, I., Nagarajan, N., and Sikîæ, M. (2017). Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.* 27, 737–746. doi: 10.1101/gr.214270.116
- Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., et al. (2014). Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* 9:e112963. doi: 10.1371/journal.pone.0112963
- Wick, R. R., Judd, L. M., and Holt, K. E. (2018). Deepbiner: demultiplexing barcoded Oxford Nanopore reads with deep convolutional neural networks. *PLoS Comput. Biol.* 14:e1006583. doi: 10.1371/journal.pcbi.1006583
- Yara, D. A., Greig, D. R., Gally, D. L., Dallman, T. J., and Jenkins, C. (2020). Comparison of shiga toxin-encoding bacteriophages in highly pathogenic strains of shiga toxin-producing *Escherichia coli* O157:H7 in the UK. *Microb. Genom.* 6:e000334. doi: 10.1099/mgen.0.000334

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Greig, Mikhail, Dallman and Jenkins. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



A Shiga Toxin-Encoding Prophage Recombination Event Confounds the Phylogenetic Relationship Between Two Isolates of *Escherichia coli* O157:H7 From the Same Patient

David R. Greig^{1,2}, Claire Jenkins^{1*} and Timothy J. Dallman^{1,2}

¹National Infection Service, Public Health England, London, United Kingdom, ²Division of Infection and Immunity, The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Easter Bush, United Kingdom

OPEN ACCESS

Edited by:

Grzegorz Węgrzyn,
University of Gdansk, Poland

Reviewed by:

Chittita Debroy,
Pennsylvania State University (PSU),
United States
Lauren Cowley,
University of Bath, United Kingdom
Atsushi Iguchi,
University of Miyazaki, Japan

*Correspondence:

Claire Jenkins
claire.jenkins1@phe.gov.uk

Specialty section:

This article was submitted to
Virology,
a section of the journal
Frontiers in Microbiology

Received: 29 July 2020

Accepted: 23 September 2020

Published: 23 October 2020

Citation:

Greig DR, Jenkins C and
Dallman TJ (2020) A Shiga
Toxin-Encoding Prophage
Recombination Event Confounds the
Phylogenetic Relationship Between
Two Isolates of *Escherichia coli*
O157:H7 From the Same Patient.
Front. Microbiol. 11:588769.
doi: 10.3389/fmicb.2020.588769

We compared genomes from multiple isolations of Shiga toxin-producing *Escherichia coli* (STEC) O157:H7 from the same patient, in cases notified to Public Health England (PHE) between 2015 and 2019. There were 261 cases where multiple isolates were sequenced from the same patient comprising 589 isolates. Serial isolates from the same patient fell within five single nucleotide polymorphisms (SNPs) of each other for 260/261 (99.6%) of the cases, indicating that there was little evidence of within host variation. The investigation into the 13 SNP discrepancy between one isolate pair revealed the cause to be a recombination event within a *stx2a*-encoding prophage resulting in the insertion/deletion of a fragment of the genome. This 50 kbp prophage fragment was homologous to a prophage in the reference genome, and the short reads from the isolate that had the 50 kbp fragment, mapped unambiguously to this region. The discrepant variants in the isolate without the 50 kbp fragment were attributed to ambiguous mapping of the short reads from other prophage regions to the 50 kbp fragment in the reference genome. Identification of such recombination events in this dataset appeared to be rare, most likely because the majority of prophage regions in the Sakai reference genome are masked during the analysis. Identification of SNPs under neutral selection, and masking recombination events, is a requirement for phylogenetic analysis used for public health surveillance, and for the detection of point source outbreaks. However, assaying the accessory genome by combining the use of short and long read technologies for public health surveillance may provide insight into how recombination events impact on the evolutionary course of STEC O157:H7.

Keywords: Shiga toxin-producing *E. coli*, genomics, Nanopore, Prophage, recombination, relatedness

INTRODUCTION

Shiga toxin-producing *Escherichia coli* (STEC) serotype O157:H7 is a zoonotic, foodborne pathogen that can cause severe gastrointestinal disease. Symptoms range from mild self-limiting diarrhea to bloody diarrhea, abdominal pain, nausea, and/or vomiting (Byrne et al., 2015). A subset of patients infected with STEC O157:H7, mainly children and the elderly, are at risk of developing hemolytic-uremic syndrome, a systemic condition associated with

renal, cardiac, and neurological complications that can be fatal (Launders et al., 2016). There are approximately 700 case reports of STEC O157:H7 in the United Kingdom each year.¹ Although case numbers are low compared to *Campylobacter* and *Salmonella*, STEC O157:H7 is regarded as a priority public health pathogen due to the potential for poor clinical outcomes. To mitigate the risks, Public Health England (PHE) operates an enhanced microbiological and epidemiological surveillance program for STEC O157:H7 (Byrne et al., 2015). All fecal specimens from hospitalized patients and from community cases reporting to primary healthcare with symptoms of gastrointestinal disease are tested for STEC O157:H7.²

All STEC O157:H7 isolated at hospital laboratories are submitted to the Gastrointestinal Bacteria Reference Unit (GBRU) at PHE, where they are sequenced to derive serotype, Shiga toxin gene (*stx*) profile, and single nucleotide polymorphism (SNP) type (Dallman et al., 2015b; Chattaway et al., 2017). Sequence similarity of pathogen genomes can be used to infer the relatedness between isolates as the fewer SNPs identified between pairs of isolates, the less time since divergence from a common ancestor (Dallman et al., 2018). SNP typing, based on hierarchical single linkage clustering of pairwise SNP distances, is used to detect outbreaks of STEC O157:H7 transmitted *via* the same vehicle and/or from the same source population (Jenkins et al., 2019).

To limit person-to-person transmission, after their symptoms have resolved, children aged five and under, food handlers and healthcare workers infected with STEC O157:H7 are required to submit fecal specimens for further testing to ensure they are no longer shedding the pathogen before returning to nursery school or work.³ Microbiological clearance testing for STEC O157:H7 is performed by hospital laboratories, and the submission of serial isolates from the same patient to GBRU is not required. However, occasionally multiple isolates from the same case are submitted to GBRU, where they are sequenced.

In the study, sequencing data from patients for whom more than one isolate of STEC O157:H7 was submitted to GBRU, were reviewed to determine the SNP difference between each isolate pair. One isolate pair from the same patient had a higher than expected SNP difference. The aim of this study was to perform long read sequencing on the two isolates from this isolate pair in order to determine the cause of this discrepancy.

MATERIALS AND METHODS

Short-Read Sequencing (Illumina HiSeq 2500) and Data Processing

Genomic DNA was extracted from cultures of STEC O157:H7 using the QIAasympyphony system (Qiagen). The sequencing library

was prepared using the Nextera XP kit (Illumina) for sequencing on the HiSeq 2500 instrument (Illumina), run with the fast protocol. FASTQ reads were processed using Trimmomatic v0.27 (Bolger et al., 2014) to remove bases with a PHRED score of <30 from the leading and trailing ends, with reads <50 bp after quality trimming discarded.

Long-Read Sequencing (Nanopore) and Data Processing

Genomic DNA was extracted and purified using the Qiagen genomic tip, midi 100/G, with minor alterations including no vigorous mixing steps (mixing performed by inversion instead) and elution into 100 µl double processed nuclease-free water (Sigma-Aldrich). Genomic DNA for each extract was quantified using a qubit and the high sensitivity (HS) dsDNA assay kit (Thermo Fisher Scientific), following the manufacturer's instructions.

Library preparation was performed using the Rapid barcoding kit SQK-RBK004 (Oxford Nanopore Technologies). The prepared libraries were loaded onto a FLO-MIN106 R9.4.1 flow cell (Oxford Nanopore Technologies) and sequenced using the MinION (Oxford Nanopore Technologies) for 24 h.

Data produced in a raw FAST5 format was basecalled and de-multiplexed using Guppy v3.2.6 using the FAST protocol (Oxford Nanopore Technologies) into FASTQ format and grouped in each samples' respective barcode. The FASTQ files were then de-multiplexed again using Deepbinner v0.2.0 (Wick et al., 2018).

Run metrics were generated using Nanoplot v1.8.1 (De Coster et al., 2018). The barcode and y-adapter from each sample's reads were trimmed, and chimeric reads split using Porechop v0.2.4 (Wick, 2017a). Finally, the trimmed reads were filtered using Filtlong v0.1.1 (Wick, 2017b) with the following parameters, min length = 1,000 bp, length_weight = 10, keep percent = 90, and target bases = 250 Mbp, to generate approximately 50x coverage of the STEC genome with the longest reads.

De novo Assembly, Polishing, Reorientation, and Annotation

Trimmed and filtered nanopore FASTQ files were assembled Flye v2.4.2 (Kolmogorov et al., 2019), using default parameters. Polishing of the assemblies was performed in a three-step process. Firstly, polishing was initiated using Nanopolish v0.11.1 (Loman et al., 2015) using both the trimmed nanopore FASTQs and FAST5s, for each respective sample accounting for methylation using the --methylation-aware = dam, dcm and --min-candidate-frequency = 0.1. The alignment was generated using Minimap2 v2.17 (Li, 2018) and Samtools v1.7 (Li et al., 2009). Secondly, the polishing was continued with Pilon v1.22 (Walker et al., 2014) using Illumina FASTQ reads as the query dataset with the use of BWA MEM v0.7.17 (Li and Durbin, 2010) and Samtools v1.7 (Li et al., 2009). Finally, Racon v1.2.1 (Vaser et al., 2017) also using BWA MEM v0.7.17 (Li and Durbin, 2010) and Samtools v1.7 (Li et al., 2009) was used with the Illumina reads to produce a final assembly for each sample.

¹<https://www.gov.uk/government/publications/escherichia-coli-e-coli-o157-annual-totals>

²<https://www.gov.uk/government/publications/smi-b-30-investigation-of-faecal-specimens-for-enteric-pathogens>

³<https://www.gov.uk/government/publications/shiga-toxin-producing-escherichia-coli-public-health-management>

As the chromosome from each assembly was circularized and closed, they were re-orientated to start at the *dnaA* gene (GenBank accession no. NC_000913) from *E. coli* K12, using the --fixstart parameter in Circlator v1.5.5 (Hunt et al., 2015). Prokka v1.13 (Seemann, 2014) with the use of a personalized database (an amino acid FASTA that included all genes annotated in the publicly available samples used in this study) was used to annotate the final assemblies.

Prophage Detection and Processing

Prophages across both samples were detected and extracted using the updated Phage Search Tool (PHASTER; Arndt et al., 2016). Prophage extraction from the genome occurred regardless of prophage size or quality and any detected prophages separated by less than 4 kbp were conjoined into a single phage using Propi v0.9.0 as described by Shaaban et al. (2016). From here the prophages were manually trimmed to remove any non-prophage genes and were again annotated using Prokka v 1.13 (Seemann, 2014) with the use of a personalized database. The output GenBank (gbk) files were modified to color genes by function.

Mash and *stx*-Encoding Prophage Phylogeny

Mash v2.2 (Ondov et al., 2016) was used to sketch (sketch length 1,000, kmer length, 21) all extracted *stx*-encoding prophages in samples 818062 and 824422 and all *stx*-encoding prophages found in the publicly available STEC genomes as described by Yara et al. (2020). The pairwise Jaccard distance between the prophages was calculated and a neighbor joining tree computed.

Variant Calling

Illumina FASTQ reads were mapped to the Sakai STEC O157 reference genome (NC_002695.1) using BWA MEM v0.7.13 (Li and Durbin, 2010) and Samtools v1.1 (Li et al., 2009). Variant positions were identified by GATK v2.6.5 UnifiedGenotyper

(McKenna et al., 2010) that passed the following parameters: >90% consensus, minimum read depth of 10, Mapping Quality (MQ) ≥ 30 , and imported into SnapperDB v0.2.5 (Dallman et al., 2018). Nanopore FASTQ reads were mapped to the Sakai STEC O157 reference genome (NC_002695.1) using Minimap2 v2.17 (Li, 2018) and Samtools v1.7 (Li et al., 2009). Methylated (5-methylcytosine) bases/positions relative to the reference genome were calculated using Nanopolish v0.11.1 (Loman et al., 2015) and masked in the alignment for each sample as described by Greig et al. (2019). The alignment for each sample was used to interrogate discrepant positions identified by SnapperDB previously.

Selection of Illumina Reads From Variant Positions and Alignment to Nanopore Assembly

Illumina reads covering the list of discrepant SNPs (Table 1) between each of the samples relative to the reference genome were identified using Samtools view v1.7 (Li et al., 2009) and read IDs using Bedtools v2.29.2 (Quinlan and Hall, 2010). These reads were deduplicated and aligned to each respective nanopore assembly and using Bedtools v2.29.2 (Quinlan and Hall, 2010) to identify where each individual read aligned to (Supplementary Material).

Data Visualization Tools

All gene diagrams were constructed using Easyfig v2.2.3 (Sullivan et al., 2011). Parsimony trees were visualized and annotated using FigTree v1.4.4 (Rambaut and Drummond, 2018). Dot plots were generated and visualized using Gepard v1.4 (Krumisiek et al., 2007).

Data Deposition

Illumina FASTQ files are available from National Centre for Biotechnology Information (NCBI) BioProject PRJNA315192 under the following SRA (sequence read archive) accession numbers: 818062; SRR10247133 and 824422; SRR10313636.

TABLE 1 | Table showing the variant positions between the two query samples (818062 and 824422) for both sequencing technologies against the reference genome (Sakai).

POS	REF	VAR	818062 Illumina	824422 Illumina	818062 ONT	824422 ONT	CDS	Locus tag	Product
2196142	G	C	G	C	G	-	L233V	ECs2204	hypothetical protein
2202081	A	G	A	G	A	G	Non coding	-	-
2202082	A	C	A	C	A	C	Non coding	-	-
2202319	C	T	C	T	C	T	Synonymous	ECs2216	putative exonuclease
2202328	C	G	C	G	C	G	Synonymous	ECs2216	putative exonuclease
2202329	C	A	C	A	C	A	H65N	ECs2216	putative exonuclease
2202334	A	G	A	G	A	G	Synonymous	ECs2216	putative exonuclease
2202338	G	A	G	A	G	A	V68I	ECs2216	putative exonuclease
2202343	G	T	G	T	G	T	Synonymous	ECs2216	putative exonuclease
2210582	A	G	A	G	A	-	Non coding	-	-
2210583	T	C	T	C	T	-	Non coding	-	-
2210594	A	G	A	G	A	-	Non coding	-	-
2237846	A	G	A	G	A	G/T	Synonymous	ECs2262	hypothetical protein

POS = position in the reference genome; REF = base in the reference genome at that position; VAR = base in the alignment at that position; A "-" refers to no reads aligned at that position (i.e., 0 depth).

Nanopore FASTQ files are available from BioProject PRJNA315192 under the following SRA accession numbers: 818062; SRR12012233 and 824422; SRR12012232.

Assemblies/draft-genomes can be found under BioProject PRJNA315192 under the following accession numbers: 818062; CP058233 (Chromosome), CP058234 (pO157) and 824422; CP058231 (Chromosome), CP058232 (pO157).

RESULTS

Analysis of Short-Read Sequencing Data From Isolates From the Same Case

Between July 2015 and December 2019, there were 261 cases where multiple isolates were sequenced from the same person comprising a total of 589 isolates (Supplementary Table). The majority of cases were associated with two isolations (215/261, 82.4%), there were 37/261 (14.2%) cases with three isolates and nine (3.4%) that were linked to between four and nine isolates. The median time between receipts of serial isolates was 6 days, with a minimum of 0 days and a maximum of 133 days. Serial isolates from the same patient fell within five SNPs of each other for 260/261 (99.6%) of the cases (Table 2). The median SNP distance between isolates from the same case was zero SNPs with a maximum of 11 SNPs (Figure 1).

One case had an isolate pair, which did not cluster into a five SNP single linkage cluster, as identified by comparing the hierarchical SNP profiles (Table 3; Dallman et al., 2018). The case was an 18-month-old female with persistent diarrhea, who had symptoms for more than 10 days prior to presenting to primary healthcare. The first isolate, designated 818062, was from a fecal specimen dated 23rd September 2019 and the second isolate, designated 824422, was from a fecal specimen

taken 10 days later on 3rd October 2019. The source of her infection was unknown. Aligned to the reference genome for SNP typing, 818062 and 824422 had 44.66x and 72.37x coverage, respectively. Further analysis of the short-read sequencing data revealed that the two isolates were 13 SNPs different, and that 11 SNPs were located in the same prophage region of the genome (Table 1) occurring in isolate 824422 with respect to the reference genome.

Analysis of the Long-Read Sequencing Data From Isolates 818062 and 834422

The genomic context of the 13 SNP differences identified in the short reads between 818062 and 824422, when compared to the Sakai reference genome, was investigated. Long read sequencing data from isolates 818062 and 824422 was assembled into two contigs for each sample. Each was identified as a single chromosome and the pO157 plasmid. The chromosome size for 818062 was 5,505,066 bp and for 824422 it was 5,457,341 bp, approximately 50 kbp different (Figure 2). The genomes of both isolates 818062 and 824422 comprised 16 prophages each, approximately 13.1 and 12.4% of each chromosome, respectively. In each isolate, three of the 16 prophages encoded *stx*; two had *stx2a* and one had *stx2c* (Figure 3). For the *stx2c*-encoding prophage the Shiga toxin-encoding bacteriophage insertion (SBI) site was *sbcb*. The SBI site of one of the *stx2a*-encoding prophages was *yecE* and the other was *rspA*.

The *stx2a*-encoding prophage inserted at *rspA* was a compound prophage, designated prophage 8. In the Sakai reference genome, Sakai prophages (SP) 11 and 12 are co-located in tandem. Prophage 8 in 81862 and 824422 has homology to SP11 and SP12 in the Sakai reference genome, with an additional *stx2a*-encoding phage inserted into SP11 component of the compound phage (Figure 4). Prophage 8 differed in size between isolates 818062 (154,371 bp; position 2,830,861–2,985,232) and 824422 (106,591 bp; position 2,830,714–2,937,305; Figure 5). The size difference was due to homologous recombination resulting in an insertion/deletion event involving a 50 kbp fragment present in prophage 8 in isolate 818062, but absent in isolate 824422.

Investigation of the Insertion/Deletion Event as the Cause of the Discrepant SNP Difference Between Isolates 818062 and 834422

To determine whether the recombination event in prophage 8 led to the SNP difference between isolates 818062 and 834422, short reads that aligned to the variant positions identified in 824422 relative to the Sakai reference genome for each sequence (shown in Table 1), were identified and mapped back each of their respective long read assemblies. There were 180 non-duplicated reads covering the variant positions relative to the Sakai reference genome for isolate 818062, and 43 non-duplicated reads for isolate 824422.

When mapped back to the long-read sequence assembly of isolate 818062, 167/180 (92.7%) of the short reads from the

TABLE 2 | Table showing total number of Shiga toxin-producing *Escherichia coli* (STEC) cases from 2015 to 2019.

	2015–16	2017	2018	2019	Total
Total number of cases	715*	563	607	514	2,399
No. of cases with serial isolation	80	60	54	67	261
No. of isolates from the same case	173	131	130	155	589
Cases with two isolates	71	51	42	51	215
Cases with three isolates	7	8	9	13	37
Cases with four isolates	1	0	1	1	3
Cases with five isolates	0	1	0	2	3
Cases with six isolates	1	0	1	0	2
Cases with nine isolates	0	0	1	0	1

*Numbers for 2016 only.

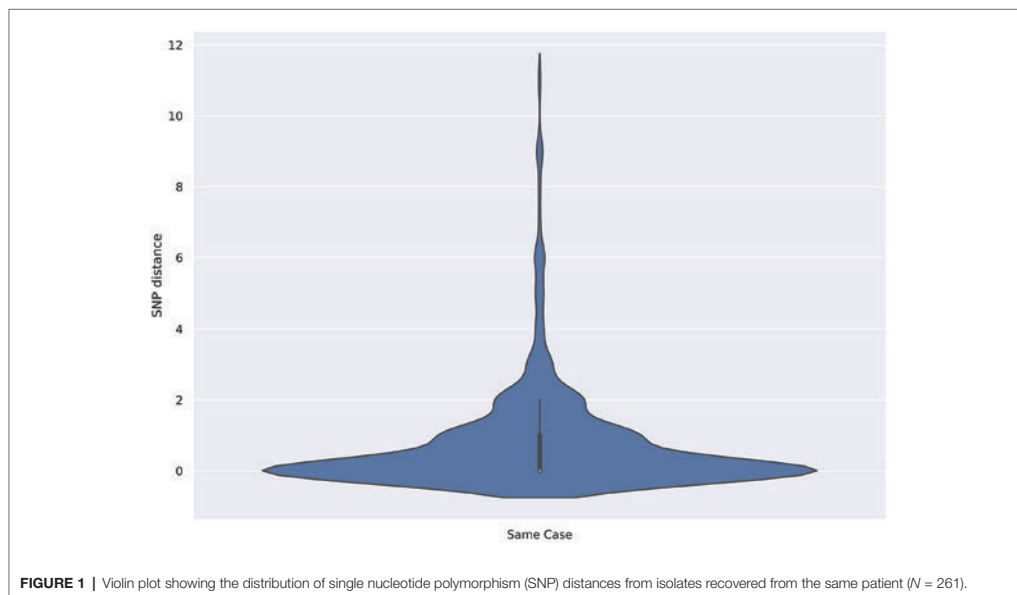


FIGURE 1 | Violin plot showing the distribution of single nucleotide polymorphism (SNP) distances from isolates recovered from the same patient ($N = 261$).

TABLE 3 | Table showing *stx* subtype, phage type, receipt date, and SNP address of samples 818062 and 824422.

Molis ID	DOB	Receipt date	Phage type	STX subtype	SNP address
818062	27/02/2018	30/09/2019	PT 34	stx2a stx2c	5.772.1448.3105.3866.5310.6343
824422	27/02/2018	14/10/2019	PT 34	stx2a stx2c	5.772.1448.3105.4913.5281.6387

Illumina data were located within the 50 kbp region of the genome (2,898,039–2,950,379 in 818062) that was not present in isolate 824422 (**Supplementary Material**). Of the remaining 13/180 (7.2%) reads, seven reads mapped within prophage 8 but outside the region where the recombination event appears to have taken place. The remaining five reads mapped to homologous regions within prophage 1 ($n = 3$ reads) and prophage 11 ($n = 2$ reads).

When aligned to the long read sequence assembly of 824422, where the 50 kbp was absent on prophage 8, 22/43 (51.1%) reads mapped back to homologous regions within prophage 1 ($n = 18$ reads), prophage 3 ($n = 1$ read), prophage 4 ($n = 1$ read), prophage 11 ($n = 1$ read), and prophage 13 ($n = 1$ read). The remaining 21/43 reads (48.9%) mapped back to prophage 8 (**Supplementary Material**).

The analysis revealed that the 50 kbp fragment present in isolate 818062, but absent in isolate 824422, also had a homologous sequence present in the Sakai reference genome (SP 11 and 12). Therefore, the short reads from the 50 kbp region on isolate 818062 mapped to the corresponding prophage region in the Sakai reference genome with fewer SNP differences than the short reads from paralogous prophage regions in isolate 824422. These short reads from homologous prophage

regions in isolate 824422 had less similarity to the homologous region in the Sakai reference genome and therefore a greater number of false positive SNP differences were detected.

To confirm that the absence of the 50 kbp prophage region in isolate 824422 was the reason for the original discrepancy, this prophage region in the Sakai reference genome was masked within the alignment, resulting in zero SNPs difference between isolates 818062 and 824422.

DISCUSSION

The relatedness between two isolate genomes can be quantified by calculating the number of SNP differences. In general, for clonal bacteria such as STEC O157:H7, the fewer polymorphisms identified between pairs of strains, the less time since divergence from a common ancestor and therefore the increased likelihood that they are from the same source population (Dallman et al., 2018; Jenkins et al., 2019). In this study, we compared SNP profiles from multiple isolations of STEC O157:H7 from the same patient, collected in response to public health guidance that requires patients in risk groups to be excluded from work or nursery school until microbiological clear. All but one isolate

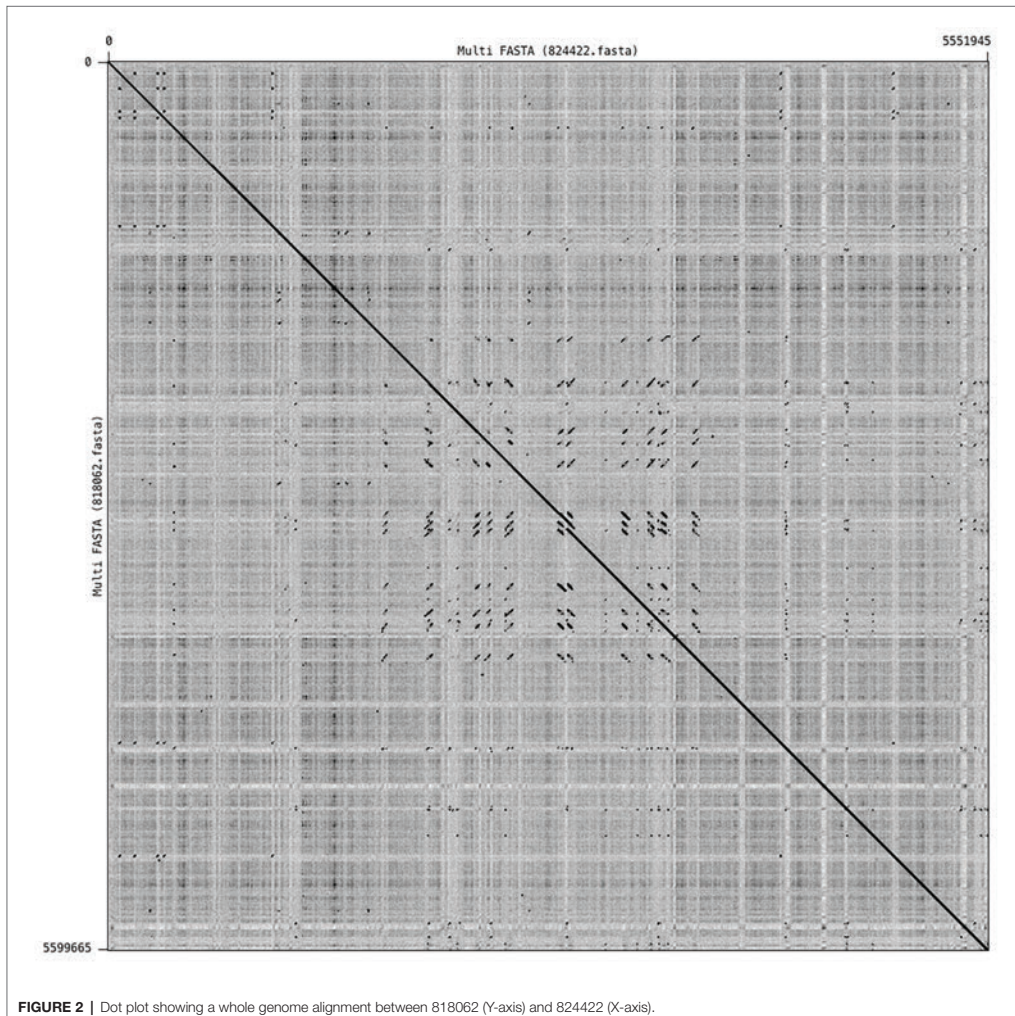


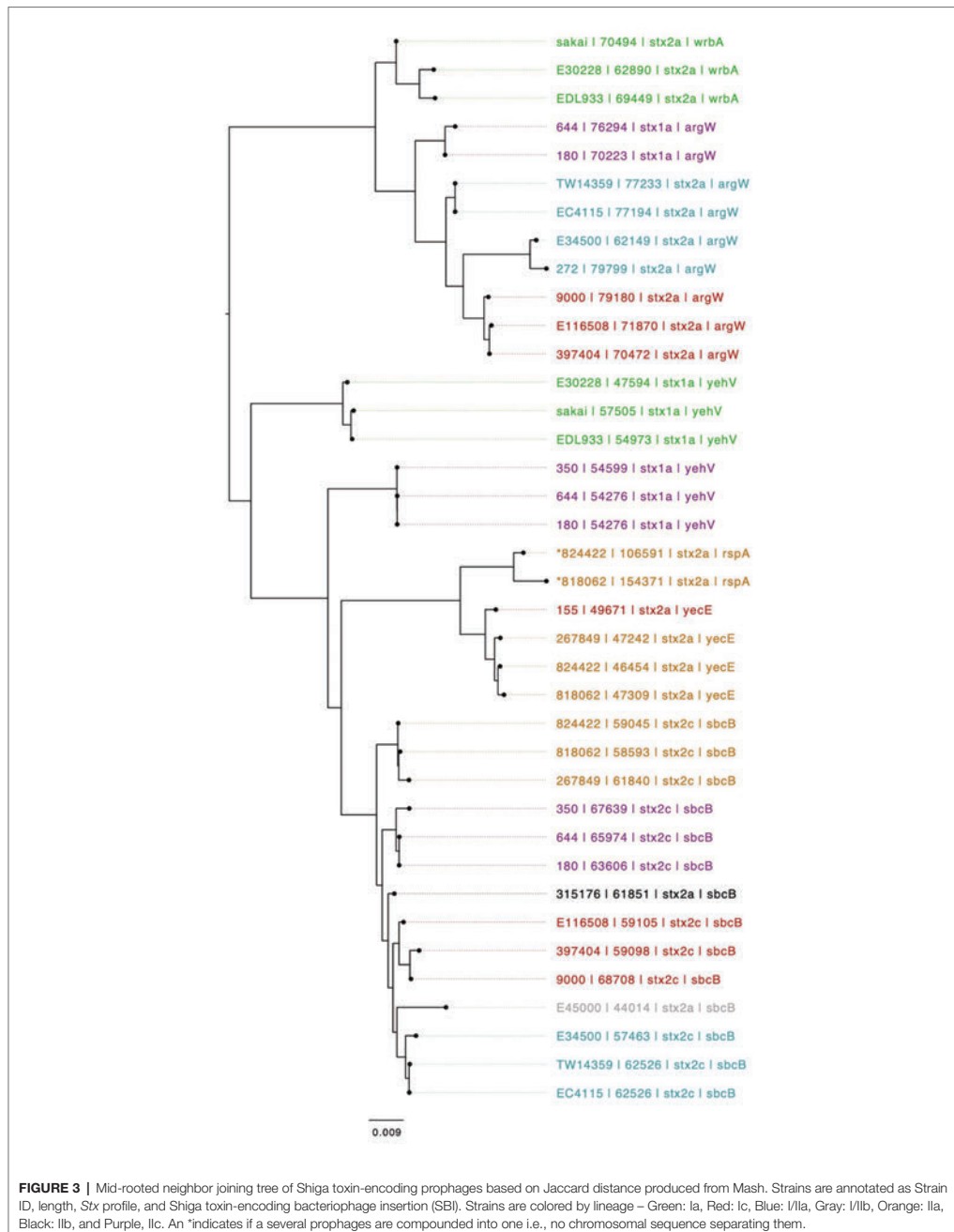
FIGURE 2 | Dot plot showing a whole genome alignment between 818062 (Y-axis) and 824422 (X-axis).

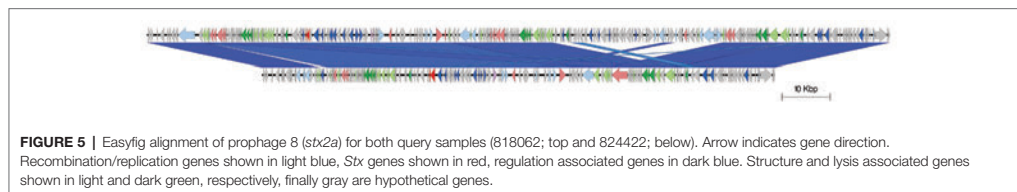
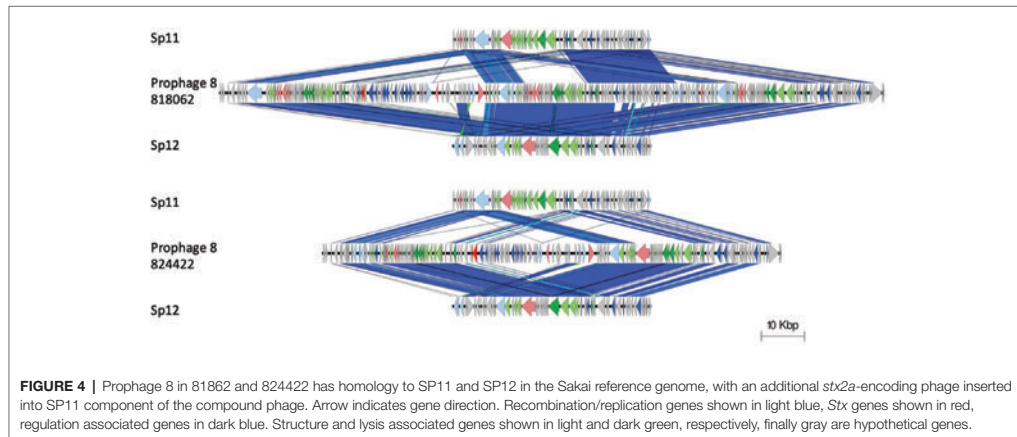
pair clustered within a five SNP single linkage cluster indicating that there is little evidence for within host variation of STEC O157:H7 within the time frame required to achieve clearance, as quantified based on the SNP typing method used at PHE (Dallman et al., 2015a).

The investigation into the 13 SNP discrepancies between one of the isolate pairs revealed the cause to be a recombination event within a *stx2a*-encoding prophage resulting in the insertion/deletion of a 50 kbp fragment of the genome. This 50 kbp prophage fragment had a homologous sequence present in the Sakai reference genome (SP11 and SP12), and the short reads

from the isolate with the 50 kbp fragment mapped unambiguously to this region in the reference genome. The variants in the isolate without the 50 kbp fragment were attributed to false mapping of the short reads to homologous regions of the reference genome (Greig et al., 2019).

Given the high percentage of prophage in STEC O157:H7 relative to other *E. coli*, and the knowledge that within host variation can occur over a short time frame, we might expect these recombination events to be detected more frequently in isolate pairs from the same patient than the initial analysis in this study suggested (Hayashi et al., 2001; Asadulghani et al., 2009;





Eppinger et al., 2011; Shaaban et al., 2016; Greig et al., 2019; Yara et al., 2020). One explanation is that the 18 prophage regions in the Sakai reference genome share similarity, particularly in genes that code for bacteriophage structures (head, tail, and portal genes) and are masked during the analysis using a reference self-mapping strategy (Dallman et al., 2018), as part of the variant calling pipeline reads simulated from the Sakai reference genome are mapped to self. Those regions of the genome where self-mapping was ambiguous, that is, where reads from multiple regions mapped to the same position, or the same reads mapped to multiple positions, are masked from any variant detection (Dallman et al., 2018).

Identification of SNPs under neutral selection, and masking recombination events, is a requirement phylogenetic analysis used for public health surveillance, and for the detection of point source outbreaks of STEC O157:H7 (Dallman et al., 2015b). However, assaying the accessory genome for further genomic characterization offers an additional level of strain of discrimination that may provide insight into the source and/or transmission of an outbreak strain (Cowley et al., 2016). As yet, we have a limited understanding of the rate of change of the recombination taking place in the STEC O157:H7 accessory genome, and the impact this has on the population structure. Combining the use of short and long read technologies for public health surveillance of STEC will improve our understanding of how microevolutionary events and large scale structural variations

in the genome contribute to persistence and survival of the pathogen in the environment, colonization and host specificity in the animal reservoir, and the emergence of clinically significant strains (Cowley et al., 2016; Shaaban et al., 2016; Yara et al., 2020).

DATA AVAILABILITY STATEMENT

Illumina FASTQ files are available from National Centre for Biotechnology Information (NCBI) BioProject PRJNA315192 under the following SRA (sequence read archive) accession numbers: 818062; SRR10247133 and 824422; SRR10313636. Nanopore FASTQ files are available from BioProject PRJNA315192 under the following SRA accession numbers: 818062; SRR12012233 and 824422; SRR12012232. Assemblies/draft-genomes can be found under BioProject PRJNA315192 under the following accession numbers: 818062; CP058233 (Chromosome), CP058234 (pO157) and 824422; CP058231 (Chromosome), CP058232 (pO157).

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent from the

participants' legal guardian/next of kin was not required to participate in this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

DG performed DNA extraction, library preparation, and Nanopore sequencing. DG performed data processing, genome assembly, correction and annotation. DG created Easyfig diagrams and performed the prophage comparison using Mash with associated scripts designed by TD. DG performed the relatedness analysis including read sub-selection and alignments. TD created violin plot. CJ and DG wrote the original manuscript. CJ, DG, and TD reviewed the manuscript. CJ and TD supervised DG. All authors contributed to the article and approved the submitted version.

REFERENCES

- Arndt, D., Grant, J. R., Marcu, A., Sajed, T., Pon, A., Liang, Y., et al. (2016). PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic Acid Res.* 44, W16–W21. doi: 10.1093/nar/gkw387
- Asadulghani, M., Ogura, Y., Ooka, T., Itoh, T., Sawaguchi, A., Iguchi, A., et al. (2009). The defective prophage pool of *Escherichia coli* O157: prophage-prophage interactions potentiate horizontal transfer of virulence determinants. *PLoS Pathog.* 5:e1000408. doi: 10.1371/journal.ppat.1000408
- Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics* 30, 2114–2120. doi: 10.1093/bioinformatics/btu170
- Byrne, L., Jenkins, C., Launders, N., Elson, R., and Adak, G. K. (2015). The epidemiology, microbiology and clinical impact of shiga toxin-producing *Escherichia coli* in England, 2009–2012. *Epidemiol. Infect.* 143, 3475–3487. doi: 10.1017/S0950268815000746
- Chattaway, M. A., Schaefer, U., Tewolde, R., Dallman, T. J., and Jenkins, C. (2017). Identification of *Escherichia coli* and *Shigella* species from whole-genome sequencing. *J. Clin. Microbiol.* 55, 616–623. doi: 10.1128/JCM.01790-16
- Cowley, L. A., Dallman, T. J., Fitzgerald, S., Irvine, N., Rooney, P. J., McAteer, S. P., et al. (2016). Short-term evolution of shiga toxin-producing *Escherichia coli* O157:H7 between two food-borne outbreaks. *Microb. Genom.* 2:e000084. doi: 10.1099/mgen.0.000084
- Dallman, T. J., Ashton, P. M., Byrne, L., Perry, N. T., Petrovska, L., Ellis, R., et al. (2015a). Applying phylogenomics to understand the emergence of shiga toxin-producing *Escherichia coli* O157:H7 strains causing severe human disease in the UK. *Microb. Genom.* 1:e000029. doi: 10.1099/mgen.0.000029
- Dallman, T., Ashton, P., Schaefer, S., Jirkin, A., Painset, A., Shaaban, S., et al. (2018). SnapperDB: a database solution for routine sequencing analysis of bacterial isolates. *Bioinformatics* 34, 3028–3029. doi: 10.1093/bioinformatics/bty212
- Dallman, T. J., Byrne, L., Ashton, P. M., Cowley, L. A., Perry, N. T., Adak, G., et al. (2015b). Whole-genome sequencing for national surveillance of shiga toxin-producing *Escherichia coli* O157. *Clin. Infect. Dis.* 61, 305–312. doi: 10.1093/cid/civ318
- De Coster, W., D'Hert, S., Schultz, D. T., Cruts, M., and Van Broeckhoven, C. (2018). NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics* 34, 2666–2669. doi: 10.1093/bioinformatics/bty149
- Eppinger, M., Mammel, M. K., Leclerc, J. E., Ravel, J., and Cebula, A. T. (2011). Genomic anatomy of *Escherichia coli* O157:H7 outbreaks. *Proc. Natl. Acad. Sci. U. S. A.* 108, 20142–20147. doi: 10.1073/pnas.1107176108
- Greig, D. R., Jenkins, C., Gharbia, S., and Dallman, T. J. (2019). Comparison of single nucleotide variants identified by illumina and oxford nanopore technologies in the context of a potential outbreak of shiga toxin-producing *Escherichia coli*. *Gigascience* 8:giz104. doi: 10.1093/gigascience/giz104
- Hayashi, T., Makino, K., Ohnishi, M., Kurokawa, K., Ishii, K., Yokoyama, K., et al. (2001). Complete genome sequence of enterohemorrhagic *Escherichia coli* O157:H7 and genomic comparison with a laboratory strain K-12. *DNA Res.* 8, 11–22. doi: 10.1093/dnares/8.1.11
- Hunt, M., Silva, N. D., Otto, T. D., Parkhill, J., Keane, J. A., and Harris, S. R. (2015). Circlator: automated circularization of genome assemblies using long sequencing reads. *Genome Biol.* 16:294. doi: 10.1186/s13059-015-0849-0
- Jenkins, C., Dallman, T. J., and Grant, K. A. (2019). Impact of whole genome sequencing on the investigation of food-borne outbreaks of shiga toxin-producing *Escherichia coli* serogroup O157:H7, England, 2013–2017. *Euro Surveill.* 24:1800346. doi: 10.2807/1560-7917.ES.2019.24.4.1800346
- Kolmogorov, M., Yuan, J., Lin, Y., and Pevzner, P. A. (2019). Assembly of long, error-prone reads using repeat graphs. *Nat. Biotechnol.* 37, 540–546. doi: 10.1038/s41587-019-0072-8
- Krumsiek, J., Arnold, R., and Rattei, T. (2007). Gepard: a rapid and sensitive tool for creating dotplots on a genome scale. *Bioinformatics* 23, 1026–1028. doi: 10.1093/bioinformatics/btm039
- Launders, N., Byrne, L., Jenkins, C., Harker, K., Charlett, A., and Adak, G. K. (2016). Disease severity of shiga toxin-producing *E. coli* O157 and factors influencing the development of typical haemolytic uraemic syndrome: a retrospective cohort study, 2009–2012. *BMJ Open* 6:e009933. doi: 10.1136/bmjopen-2015-009933
- Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 34, 3094–3100. doi: 10.1093/bioinformatics/bty191
- Li, H., and Durbin, R. (2010). Fast and accurate long-read alignment with burrows–wheeler transform. *Bioinformatics* 26, 589–595. doi: 10.1093/bioinformatics/btp324
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi: 10.1093/bioinformatics/btp352
- Loman, N. J., Quick, J., and Simpson, J. T. (2015). A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nat. Methods* 12, 733–735. doi: 10.1038/nmeth.3444
- McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernysky, A., et al. (2010). The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20, 1297–1303. doi: 10.1101/gr.107524.110
- Ondov, B. D., Treangen, T. J., Melsted, P., Mallonee, A. B., Bergman, N. H., Koren, S., et al. (2016). Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.* 17:132. doi: 10.1186/s13059-016-0997-x
- Quinlan, A. R., and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26, 841–842. doi: 10.1093/bioinformatics/btq033
- Rambaut, A., and Drummond, A. J. (2018). Available at: <http://tree.bio.ed.ac.uk/software/figtree/>

FUNDING

The research was part funded by the National Institute for Health Research (NIHR) Health Protection Research Unit in Gastrointestinal Infections at the University of Liverpool (UK), in partnership with PHE, in collaboration with the University of East Anglia (UK), the University of Oxford (UK), and the Quadram Institute (UK). CJ, TD, and DG are based at PHE.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fmicb.2020.588769/full#supplementary-material>

- Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 30, 2068–2069. doi: 10.1093/bioinformatics/btu153
- Shaaban, S., Cowley, L., McAteer, S. P., Jenkins, C., Dallman, T. J., Bono, J. L., et al. (2016). Evolution of a zoonotic pathogen: investigating prophage diversity in enterohaemorrhagic *Escherichia coli* O157 by long-read sequencing. *Microb. Genom.* 2:e000096. doi: 10.1099/mgen.0.000096
- Sullivan, M. J., Petty, N. K., and Beatson, S. A. (2011). Easyfig: a genome comparison visualizer. *Bioinformatics* 27, 1009–1010. doi: 10.1093/bioinformatics/btr039
- Vaser, R., Sović, I., Nagarajan, N., and Šikić, M. (2017). Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.* 27, 737–746. doi: 10.1101/gr.214270.116
- Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., et al. (2014). Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* 9:e112963. doi: 10.1371/journal.pone.0112963
- Wick, R. R. (2017a). Porechop. Available at: <https://github.com/rrwick/Porechop>
- Wick, R. R. (2017b). Filtlong. Available at: <https://github.com/rrwick/Filtlong>
- Wick, R. R., Judd, L. M., and Holt, K. E. (2018). Deepbiner: demultiplexing barcoded oxford nanopore reads with deep convolutional neural networks. *PLoS Comput. Biol.* 14:e1006583. doi: 10.1371/journal.pcbi.1006583
- Yara, D. A., Greig, D. R., Gally, D. L., Dallman, T. J., and Jenkins, C. (2020). Comparison of shiga toxin-encoding bacteriophages in highly pathogenic strains of shiga toxin-producing *Escherichia coli* O157:H7 in the UK. *Microb. Genom.* 6:e000334. doi: 10.1099/mgen.0.000334

Disclaimer: The views expressed are those of the authors and not necessarily those of the National Health Service, the NIHR, the Department of Health nor PHE.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Greig, Jenkins and Dallman. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Analysis of a small outbreak of Shiga toxin-producing *Escherichia coli* O157:H7 using long-read sequencing

David R. Greig^{1,2}, Claire Jenkins^{1,*}, Saheer E. Gharbia¹ and Timothy J. Dallman^{1,2}

Abstract

Compared to short-read sequencing data, long-read sequencing facilitates single contiguous *de novo* assemblies and characterization of the prophage region of the genome. Here, we describe our methodological approach to using Oxford Nanopore Technology (ONT) sequencing data to quantify genetic relatedness and to look for microevolutionary events in the core and accessory genomes to assess the within-outbreak variation of four genetically and epidemiologically linked isolates. Analysis of both Illumina and ONT sequencing data detected one SNP between the four sequences of the outbreak isolates. The variant calling procedure highlighted the importance of masking homologous sequences in the reference genome regardless of the sequencing technology used. Variant calling also highlighted the systemic errors in ONT base-calling and ambiguous mapping of Illumina reads that results in variations in the genetic distance when comparing one technology to the other. The prophage component of the outbreak strain was analysed, and nine of the 16 prophages showed some similarity to the prophage in the Sakai reference genome, including the *stx2a*-encoding phage. Prophage comparison between the outbreak isolates identified minor genome rearrangements in one of the isolates, including an inversion and a deletion event. The ability to characterize the accessory genome in this way is the first step to understanding the significance of these microevolutionary events and their impact on the evolutionary history, virulence and potentially the likely source and transmission of this zoonotic, foodborne pathogen.

DATA SUMMARY

All FASTQ files and assemblies were submitted to the National Centre for Biotechnology Information (NCBI). All data can be found under BioProject: PRJNA315192 - <https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA315192>. Strain-specific details can be found in Methods under data deposition.

INTRODUCTION

Shiga toxin-producing *Escherichia coli* (STEC) O157:H7 is a human, gastrointestinal pathogen that colonizes the gut of healthy ruminants, particularly cattle and sheep. Symptoms in humans range from mild diarrhoea to include abdominal cramps, vomiting and severe bloody diarrhoea. In 5–15% of cases, the infection can lead to the development of haemolytic uremic syndrome (HUS), a severe multi-system syndrome

[1], that can be fatal, particularly in young children and the elderly. STEC O157:H7 has a very low infectious dose (10–100 organisms) and transmission to humans occurs through consumption of contaminated food or water, direct or indirect contact with animals or their environment and through person-to-person spread [1].

In 2015, Public Health England (PHE) implemented high-throughput, real-time sequencing for the surveillance of gastrointestinal pathogens, including STEC O157:H7. The detection of SNPs by mapping short reads to a single reference genome is used to identify linked cases and outbreaks of infectious disease. High-quality SNPs are identified based on validated thresholds of mapping quality, mapping depth, and variant ratio. SNPs that do not meet these criteria, positions that have no aligned reads, or invariant positions with depth or mapping quality less than the specified thresholds are

Received 09 October 2020; Accepted 15 February 2021; Published 08 March 2021

Author affiliations: ¹National Infection Service, Public Health England, London, NW9 5EQ, UK; ²Division of Infection and Immunity, The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Easter Bush, EH25 9RG, UK.

*Correspondence: Claire Jenkins, Claire.Jenkins1@phe.gov.uk

Keywords: Bacteriophage; *Escherichia coli* O157:H7; Illumina; Nanopore; Shiga toxin; Whole Genome sequencing.

Abbreviations: GATK, genome analysis toolkit; HUS, haemolytic uraemic syndrome; LEE, locus of enterocyte effacement; MQ, mapping quality; NCBI, national centre for biotechnology information; ONT, Oxford Nanopore Technology; PHE, Public Health England; PLE, prophage-like element; PT, phage type; SBI, Shiga toxin bacteriophage insertion; STEC, Shiga toxin-producing *Escherichia coli*; WGS, whole genome sequencing.

Data statement: All supporting data, code and protocols have been provided within the article or through supplementary data files.

000545 © 2021 The Authors



This is an open-access article distributed under the terms of the Creative Commons Attribution NonCommercial License. This article was made open access via a Publish and Read agreement between the Microbiology Society and the corresponding author's institution.

termed 'ignored positions'. Consequently, a high proportion of repetitive and homologous features, including prophage, are masked from the analysis, and little is known about the variation in prophage content of STEC O157:H7 genomes.

STEC O157:H7 has a large accessory genome, with approximately 10–15% of the genome comprised of prophage [2, 3]. Furthermore, the defining characteristic of the STEC group, the Shiga toxin genes (*stx*) are bacteriophage encoded [4]. Therefore, analysis of prophage content, loss and acquisition of bacteriophage and structural rearrangements within prophage regions contributes to our understanding of the evolutionary history, virulence and potentially the likely source and transmission of this zoonotic, foodborne pathogen. Long-read sequencing technologies, such as Oxford Nanopore Technology (ONT) have been shown to achieve improved *de novo* assemblies and facilitate more complete characterization of the accessory genome [5, 6] including prophage regions [6].

In August 2017, a cluster of four cases (A–D) infected with genetically related strains of STEC O157:H7 was identified by the national Gastrointestinal Infections Department at Public Health England (https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/765498/STEC_O157_PT21.28_Outbreak_Report.pdf). All four cases were identified as STEC O157:H7 phage type 21/28 harbouring *stx* subtypes, *stx2a* and *stx2c*, and belonging to sub-lineage 1c [7]. The strains possessed the *stx2a* toxin subtype, known to be associated with more severe disease and HUS and, despite the small numbers of cases, a multi-agency investigation was undertaken. Handling raw pet food, specifically tripe (the edible lining of the stomach of cattle and sheep), was identified as the cause of the outbreak. The SNP-type profile derived from the short-read Illumina sequencing data for three cases were identical and one isolate (from case B) differed by one SNP from the other isolates (Fig. 1). We describe our methodological approach to the analysis of ONT sequencing data to further quantify genetic relatedness and to look for microevolutionary events in the core and accessory genomes to assess the within-outbreak variation of four genetically and epidemiologically linked isolates.

METHODS

Short-read sequencing on the Illumina platform and core SNPs analysis

Genomic DNA was extracted from cultures of STEC O157:H7 using the Qiagen Qiasymphony (Qiagen, Hilden, Germany). The sequencing library was prepared using the Nextera XP kit (Illumina, San Diego, USA) for sequencing on the Illumina HiSeq 2500 (Illumina, San Diego, USA) instrument run with the fast protocol. High-quality trimmed (leading and trailing trimming at <Q30 using Trimmomatic v0.27 [8]). Illumina reads (read length 80–100 bp) were mapped to the STEC O157:H7 reference genome Sakai (GenBank accession BA000007) using BWA-MEM v0.7.13 [9]. The Sakai STEC O157:H7 reference genome (BA000007) contains 18 prophages of which two are *Stx*-encoding (*stx1a* and *stx2a*)

Impact Statement

The use of short-read sequencing data for surveillance of gastrointestinal pathogens is well established, and the added value provided by this approach has been well documented. Here, we begin to explore how supplementing short-read sequencing data with long-read sequencing data (Oxford Nanopore Technology) can add value to public health surveillance of STEC, including outbreak detection and investigation. We describe our methodological approach to the analysis of the accessory genomes of four temporally related cluster isolates of STEC O157:H7. The comparison of the ONT sequencing data with the Illumina sequencing data confirmed the close genetic relatedness of the four outbreak isolates. Although between the outbreak strains the prophage content was stable, minor structural alterations were observed in two prophages in one of the isolates. Long-read sequencing data provides an opportunity to explore the accessory genome, and to better understand the significance of these microevolutionary events.

and six prophage like-regions including the locus of enterocyte effacement [10]. SNPs were identified using GATKv2.6 in unified genotyper mode [11]. Core-genome positions that had a high-quality SNP (>90% consensus, minimum depth 10×, MQ ≥30) in at least one isolate were extracted for further analysis. Genomes were compared to the sequences held in the PHE STEC O157:H7 WGS database, (SnapperDB v0.2.5. STEC O157:H7) and isolates with five SNP differences or less within their core genome were considered closely related and likely to have an epidemiological link [7, 12].

Long read sequencing using ONT and data processing

Genomic DNA was extracted and purified using two methods. The first was the Promega Wizard Genomic DNA Purification Kit (Promega, Madison, USA) with minor alterations including doubled incubation times, no vigorous mixing steps (performed by inversion) and elution into 50 µl of double processed nuclease free water (Sigma-Aldrich, St. Louis, USA). The second method was the Revoluten Fire Monkey DNA extraction kit (Revoluten, Glossop, UK) to the manufacturer's instructions. DNA was quantified using a Qubit and the HS (High sensitivity) dsDNA Assay Kit (ThermoFisher Scientific, Waltham, USA) to the manufacturer's instructions. Library preparation was performed using the Native Barcoding kit (SQK-LSK108 and EXP-NBD103) (Oxford Nanopore Technologies, Oxford, UK). The prepared library was loaded on a FLO-MIN106 R9.4.1 flow cell (Oxford Nanopore Technologies, Oxford, UK) and sequenced using the MinION for 48 h.

Data was produced in a raw FAST5 format was base-called and de-multiplexed using Guppy V3.2.6 (Oxford Nanopore

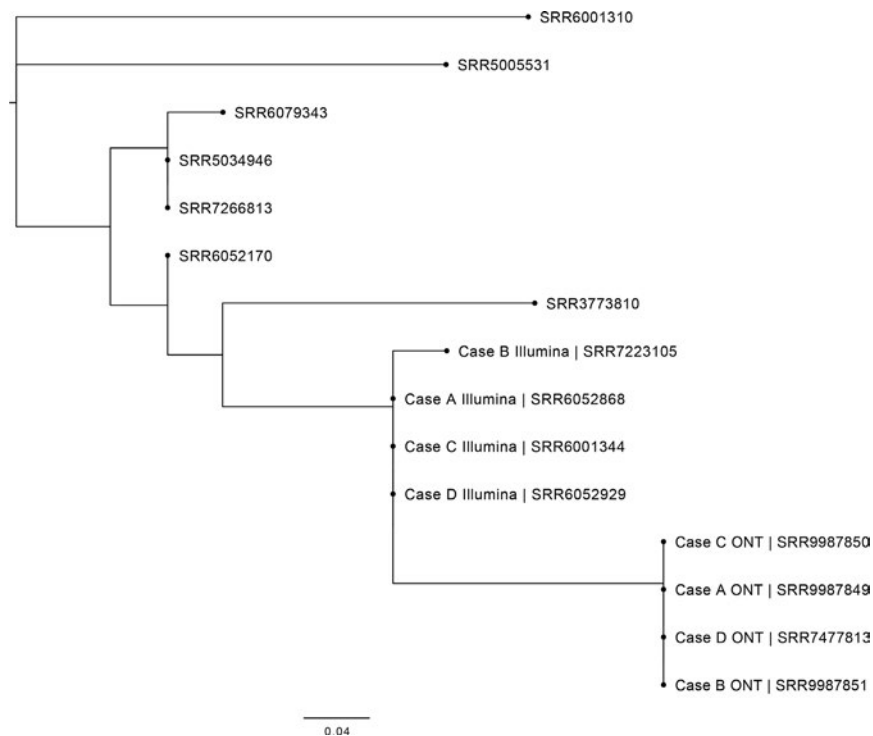


Fig. 1. Maximum-likelihood phylogenetic tree of the four outbreak samples within this outbreak and the seven most closely related isolates in the PHE archive to show context. The four outbreak cases (A–D) sequenced by both Illumina and Nanopore technologies are shown.

Technologies) into FASTQ format and grouped in each samples' respective barcode. Samples were re-demultiplexed using Deepbiner v0.2.0 [13]. Run metrics were generated using Nanoplot v1.8.1 [14]. The barcode and y-adapter from each sample's reads were trimmed, and chimeric reads split using Porechop v0.2.1 [15]. Finally, the trimmed reads were filtered using Filtlong v0.1.1 [16] with the following parameters; min_length=1000, keep_percent=90 and target_bases=550 Mbp, to generate approximately 100× coverage of the STEC genome with the longest and highest quality reads.

De novo assembly, polishing, reorientation and annotation

Trimmed and filtered ONT FASTQ files were assembled using Flye v2.6 [17]. The assembly for each sample that had the highest N50 and lowest number of contigs with the assembly size (between 5.3–6.0 Mbp) were taken forward. Polishing of the assemblies was performed in a three-step process firstly, using Nanopolish v0.11.1 [18] using both the trimmed ONT FASTQs and FAST5s for each respective sample accounting for

methylation using the --methylation-aware=dcn,dam, --min-candidate-depth=10 and --min-candidate-frequency=0.1. Secondly, Pilon v1.22 [19] with --minmq=0, --minqual=0 and --mindepth=0.05 set. Illumina FASTQ reads were used as the query dataset with the use of BWA v0.7.17 [9] and Samtools v1.7 [20]. Finally, Racon v1.2.1 [21] (--error-threshold=0.3 and --quality-threshold=10) also using BWA v0.7.17 [9] was used with the Illumina reads to produce a final assembly for each sample. As all assemblies were circularized and closed, they were reoriented to start at the *dnaA* gene (NC_000913) from *E. coli* K12, using the --fixstart parameter in circlator v1.5.5 [22]. Prokka v1.13 [23] with the use of a personalized database (https://github.com/gingerdave269/prophage_DB) was used to annotate the final assemblies.

Prophage detection, excision and processing

Prophages across all samples were detected using the Phage Search Tool (PHASTER) [24]. Prophage sequences were extracted from each samples' chromosome and this occurred regardless of prophage size or quality. Any detected

Table 1. Total number of SNPs for each sample relative to the each other for both Illumina and Nanopore technologies

Sample	Case A ONT	Case B ONT	Case C ONT	Case D ONT	Case A Illumina	Case B Illumina	Case C Illumina	Case D Illumina
Case A ONT	/	0	0	0	5	6	5	5
Case B ONT	0	/	0	0	5	6	5	5
Case C ONT	0	0	/	0	5	6	5	5
Case D ONT	0	0	0	/	5	6	5	5
Case A Illumina	5	5	5	5	/	1	0	0
Case B Illumina	6	6	6	6	1	/	0	0
Case C Illumina	5	5	5	5	0	0	/	0
Case D Illumina	5	5	5	5	0	0	0	/

Table 2. Table showing the position of the variants between the two sequencing technologies for all four samples

Position in reference genome	Base in reference genome	Base in Illumina data	Base in nanopore data	
270595	C	A	C	False positive by Illumina
379516	A	G	A	False negative by Nanopore
2033176	T	G	T	False negative by Nanopore
4709195	A	A	G	False positive by Nanopore
4901209	A	A	G	False positive by Nanopore

prophages separated by less than 4 kbp were conjoined into a single phage using Propi v0.0.1 as described in Shaaban *et al.* [25]. Prophages were re-annotated using Prokka v 1.13 [23]. Prophages were compared using Easyfig v2.2.5 [26].

Mash and prophage comparison

Mash v2.2 [27] was used to sketch (sketch length 1000, kmer length, 21) all extracted prophages in the samples sequenced in this study and all prophages found in the Sakai STEC reference genome (BA000007). The pairwise Jaccard distance between the prophages was calculated and a neighbour-joining tree computed and visualised using FigTree v1.4.4.

Variant calling and phylogenetic tree construction

For reference-based variant calling both Illumina and ONT FASTQ reads were mapped to the Sakai STEC O157 reference genome (BA000007) using BWA v0.7.3 and minimap2 v2.2, respectively [28]. VCFs were produced using GATK v2.6.5 UnifiedGenotyper [11]. Core-genome positions that had a high-quality SNP ($>90\%$ for Illumina) [$>80\%$ for ONT] consensus, minimum depth 10x, MQ ≥ 30 in at least one isolate were extracted for further analysis. Any variants called at positions that were within the known prophages in Sakai were masked from further analyses. 5-methylcytosine positions were identified using Nanopolish V0.11.1 [18] and methylated positions were then masked from the ONT VCFs as described in Greig *et al.* [29]. Masking the prophage regions and relative methylated positions of the reference genome leads to an 81.0% core genome to compare for both

Illumina and Nanopore data for the four outbreak samples. The maximum-likelihood phylogenetic tree was constructed by RAxML v8.1.17 [30] using an alignment generated from SnapperDB [12] that recombination had been accounting for by Gubbins v2.00 [31]. Visualization of the phylogenetic tree was performed using FigTree v1.4.4 (Fig. 1). To detect false positive/negative SNPs called by Illumina reads, discrepant variant positions between Illumina and Nanopore relative to the reference genome were extracted. Those variants that were called in paralogous sequences that also had a lower-than-average mapping quality were then masked in the alignment. To be included in the masking process the false called variant must be present in all the alignment of all samples in the study.

Data deposition

All FASTQ and assemblies were submitted to the National Centre for Biotechnology Information (NCBI). Illumina FASTQ accessions: case A: SRR6052868, case B: SRR7223105, case C: SRR6001344 and case D: SRR6052929. Nanopore FASTQ accessions: case A: SRR9987849, case B: SRR9987851, case C: SRR9987850 and case D: SRR7477813. All FASTQs can be found under BioProject: PRJNA315192. Assembly accessions (chromosome and plasmid): case A: CP043011 and CP043012, case B: CP043015 and CP043016, case C: CP043019 and CP043020, case D: CP043025 and CP043023. All FASTQs can be found under BioProject: PRJNA315192.

RESULTS AND DISCUSSION

Assemblies generated by long-read sequencing data and variant calling

The assemblies for each isolate were resolved into two contigs comprising the chromosome and pO157 plasmid. The chromosome length (plasmid sizes and type) was 5486665 bp (91449 bp IncFIB), 5487004 bp (91445 bp IncFIB and 57390 bp IncI2), 5486935 bp (91445 bp IncFIB and 17157 bp unknown type) and 5424337 bp (91443 bp IncFIB) and for cases A to D respectively. Sample B contained an extra 57kbp IncI2 plasmid.

In total, variant calling using GATK and SnapperDB identified between zero and six SNPs when comparing each sample with both sequencing technologies to one another (Table 1). Concerning the Illumina generated sequence data, there was only a single SNP detected between each sample. This SNP was located in sample B at position 2578517 relative to the reference genome within a gene that encodes a proton conductor (Table 1). For the ONT generated sequence data, there were no SNPs detected between all four samples relative to the reference genome (Table 2). There were five SNPs that differed between ONT and Illumina datasets relative to the reference genome. Of the five discrepant SNPs, one was called as a variant in the Illumina data and four were called as variants in the ONT data (Table 2).

All five of the discrepant SNPs between both technologies were false positive or false negative calls. Ambiguous mapping of short-read Illumina sequences to paralogous sequences in the reference genome leads to the introduction of a single false positive SNP. The remaining four false positive or false negative SNPs were generated from a known systemic error during the base-calling process of homopolymer regions in ONT sequencing, resulting in small single or double base insertions within the reads [29, 32]. Correction for these systemic errors and the ambiguous mapping error confirmed a single SNP in sample B in the Illumina dataset and all other samples (and technologies) had no SNPs different from each other.

The comparison of the ONT sequencing data with the Illumina sequencing data confirmed the close genetic relatedness of the four outbreak isolates, as only one additional SNP was identified between the outbreak strain genomes. This comparison highlighted the limitations associated with each technology, specifically the base-calling errors related to homopolymer detection observed in ONT data and the importance of masking of homologous and paralogous regions in the Illumina data.

Analysis of prophage content of the outbreak isolates and comparison with the Sakai reference genome

Of 16 prophage regions, 15 were shared between the four outbreak isolates (Tables 3 and 4, Fig. 2), with sample D containing an extra prophage (Table 3). Prophage size ranged from 8 to 145 kbp. Seven of the 16 prophages showed similarity to prophages in the Sakai reference genome; all seven

prophages had 98–100% nucleotide identity and coverage (Table 3 and Fig. 2). Prophages 3, 4, 5, 7, 10, 14 and 15 in the outbreak isolates matched >98% similarity and coverage to Sakai prophages (Sp3, Sp4, Sp6, Sp8, Sp14, Sp16 and Sp17, respectively), and shared the same bacteriophage insertion (SBI) site (*ybhC*, *yccA*, *potC*, *icd*, *serU*, *argW* and *ssrA*, respectively). Case D had an extra prophage compared to the other three samples designated prophage A (Figs 2 and 3, Table 3). This prophage was 28 kbp in length and integrated at the *hipA* gene.

There were nine prophages in the outbreak isolates and 11 in Sakai that were <50% homologous or do not match at all. Five prophages in the outbreak isolates share the same SBI site with prophages in the Sakai reference genome; prophage 6 shares *phoQ* with Sp6, prophage 12 shares *yehV* with Sp15. There appears to be a homologous recombination event in between prophages 8 and 9 relative to Sp11 and 12. Prophage 2 labelled as a compound prophage shares *thrW* where Sp1 and Sp2 are located. The sites of prophages 1, 11 and 13 located at *lexA*, *tnpA* and *argW*, respectively, in the outbreak strain are vacant in the Sakai reference genome whereas the sites for Sp5, Sp9, Sp10, Sp13 and Sp18 located at *wrbA*, *yciD*, *ydaO*, *leuZ* and a sorbitol operon, respectively in Sakai are vacant in the outbreak samples.

In the outbreak strain, *stx2a* and *stx2c* were encoded on prophages 11 and 13, inserted at *argW* and *tnpA*, respectively. For the *stx2c* encoding prophage the known SBI site, *sbvB*, as previously described for a PT21/28 STEC [15, 29, 33] has been split by a short 2.7 kbp insertion sequence (IS629) hence the designation *tnpA* which encodes a transposase.

The *stx2a*-encoding prophage detected in the strain described in this study had ~30% coverage but greater than 97% nucleotide similarity with Sp5, which is the Sakai *stx2a*-encoding prophage. The regions of high similarity included the *stx* encoding genes, Q region, *nin* region, DNA replication, origin and general recombination and the prophage structural regions differed including head, tail and tail fibres/tip regions. The *stx2c*-encoding prophage was not present in the Sakai reference strain and so no comparison was possible. Unlike Sakai, the samples sequenced in this study did not contain a *stx1a*-encoding prophage however, Sp15 which is a *stx1a*-encoding prophage was structurally similar to that of prophage 11 and shared the same SBI site, *yehV* (Table 3, Fig. 3).

The strain of STEC O157:H7 linked to the tripe outbreak sequenced in this study and the Sakai reference strain [34] belonging to two different sub-lineages, sub-lineages 1c and 1a, respectively, were isolated in geographically distinct regions, 20 years apart. The prophage commonality shared between the two strains indicates some stability of the non-*stx*-encoding prophage content over time and space (Fig. 3). In contrast, the dynamic nature of the *stx*-encoding phage is well documented, and variation of *stx* profiles in strains belonging to the same lineage that are globally distributed but also in closely related strains at the local level has been described [7, 34, 35]. Previous studies charting the evolutionary history

Table 3. The size, position and integration sites of all 16 prophages within all samples in this study, also included is the percentage nucleotide similarity to any Sakai prophages

Prophage detected	Gene 5' to prophage	Gene 3' to prophage	Size (bp) in sample A	Size (bp) in sample B	Size (bp) in sample C	Size (bp) in sample D	(% similarity /% coverage) to Sakai prophages	Position in sample A	Position in sample B	Position in sample C	Position in sample D
1	<i>lexA</i>	<i>aphA</i>	22035	22035	22034	22036	/	403708-425743	403729-425764	403712-425746	403708-425744
2*	<i>tRNA-Thr (cgt)†</i>	<i>prgR</i>	26826	26823	26823	26821	/	1101413-1128239	1101440-1128263	1101421-1128244	1101414-1128235
3	<i>ydhC†</i>	<i>yhbB</i>	38095	38095	38096	38095	Sp3 (99%/100%)	1697277-1735372	1697420-1735515	1697397-1735493	1697388-1735483
4	<i>yccA†</i>	<i>tRNA-Ser (tga)</i>	48257	48277	48265	48263	Sp4 (99%/100%)	1967370-2015627	1967513-2015790	1967491-2015756	1967481-2015744
5	<i>potC†</i>	<i>palB</i>	47889	47679	47879	47896	Sp6 (99%/100%)	2283891-2331780	2284054-2331733	2284019-2331898	2284012-2331908
6	<i>roxA†</i>	<i>phoQ</i>	10436	10436	10436	10436	/	2337022-2347458	2336975-2347411	2337140-2347576	2337150-2347586
7	<i>icd†</i>	<i>cadB</i>	40992	40991	40992	40991	Sp8 (99%/100%)	2356230-2397222	2356183-2397174	2356348-2397340	2356358-2397349
8*	<i>ompW†</i>	<i>rspR</i>	145151	144981	145040	53636	/	2495566-2640717	2495518-2640499	2495685-2640725	2495693-2549329
A	<i>hipA</i>	<i>ydaP</i>	/	/	/	28513	/	/	/	/	2589222-2617735
9	<i>trpA</i>	<i>rspA</i>	59339	59340	59322	59863	/	2918047-2977386	2917834-2977174	2918061-2977383	2855169-2915032
10	<i>yadB</i>	<i>tRNA-Ser (cga)†</i>	43681	43694	43684	43690	Sp14 (99%/98%)	3370042-3413723	3369847-3413541	3370056-3413740	3307170-3350860
11 (<i>stx2c</i>)	<i>yxeA</i>	<i>trpA†</i>	55650	56910	55652	55655	/	3457661-3513311	3456222-3513132	3457677-3513329	3394797-3450452
12	<i>ydhV</i>	<i>ydhY†</i>	45720	45725	45732	45719	/	3654766-3700486	3654584-3700309	3654782-3700514	3591904-3637623
13 (<i>stx2a</i>)	<i>yjdC</i>	<i>argW†</i>	60239	60237	60239	60242	/	3948563-4008802	3948386-4008623	3948591-4008830	3885700-3945942
14	<i>argW†</i>	<i>lacY</i>	8233	8233	8233	8233	Sp16 (100%/100%)	4009234-4017467	4009055-4017288	4009262-4017495	3946374-3954607
15	<i>sraA</i>	<i>alpA†</i>	22107	22103	22098	22106	Sp17 (99%/99%)	4294770-4316877	4294584-4316687	4294801-4316899	4231915-4254021

*Refers to prophages that appear to be compound prophages (i.e. two or more prophages that are sequential, with intact integrase genes).

†Refers to the end in which the integrase gene (*intA*) is located.

Table 4. The size, position and integration sites of any prophage-like elements within all samples in this study. Also showing the percentage nucleotide similarity and coverage to the prophage-like elements found in Sakai

Prophage-like element	Gene 5'	Gene 3'	Size (bp) in sample A	Size (bp) in sample B	Size (bp) in sample C	Size (bp) in sample D	(% similarity / %coverage) to Sakai SpLE's	Position in sample A	Position in sample B	Position in sample C	Position in sample D
PLE1	tRNA-Leu	int	9530	9530	9530	9530	SpLE5 (99%/97%)	650711–660241	650730–660260	650601–660131	650715–660245
PLE2	int	nanS	34465	34981	34465	34465	SpLE6 (99%/100%)	660260–694725	659763–694744	660150–694615	660264–694729
PLE3	ycdU	tRNA-Ser	85906	85906	85906	85911	SpLE1 (99%/100%)	2112819–2198725	2113064–2198970	2112913–2198819	2113028–2198939
PLE4	cobU	ycdB	15044	15045	15044	15046	SpLE2 (99%/100%)	3439339–3454383	3439883–3454928	3439719–3454763	3377102–3392148
PLE5	tRNA-Phe	ptbB	23334	23333	23333	23332	SpLE3 (99%/100%)	4664210–4687544	4664345–4687678	4664588–4687921	4601991–4625323
PLE6	selC	ycdL	50390	50389	50390	50390	SpLE4 (99%/79%)	5392598–5442988	5392972–5443361	5393215–5443605	5330617–5381007

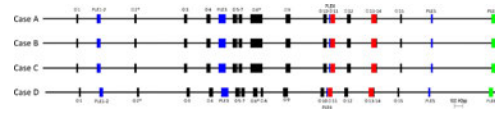


Fig. 2. The relative positions of the 15 prophages and six prophage-like elements within the chromosomes of samples A–C, 16 prophages within sample D (order descending). Red prophages indicate *stx* gene encoding prophages. Blue indicates prophage like elements. Green indicates the locus of enterocyte effacement (LEE).

of STEC O157:H7 propose the lineage I progenitor strain has *stx2c* only [7]. At some point during its evolutionary history, the Sakai outbreak strain appears to have lost the *stx2c*-encoding prophage and acquired a *stx1a*-encoding (which is similar to the *stx*-negative prophage 11 in this study) and a *stx2a*-encoding prophage, although the order of these events is unclear. The acquisition of *stx2a*-encoding prophages by sub-lineage 1 c in the UK approximately 25–30 years ago is well described and resulted in the change in PT from PT32 to PT21/28 [7, 25, 34]. The *stx2a*-encoding prophage (prophage 12) and Sp5 (Sakai's *stx2a*-encoding prophage) share only 40% of hashes via mash and both have different SBI sites (*argW* and *wrbA*, respectively).

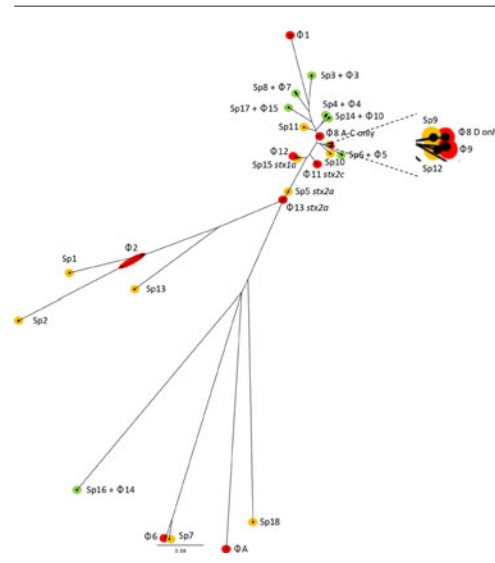


Fig. 3. Mid-rooted neighbour-joining tree of Jaccard distances showing prophages from cases A–D with prophages from BA000007 (Sakai). Grouped by green; prophages shared in cases A–D and Sakai, yellow; Sakai only and red; prophages unique to cases A–D.

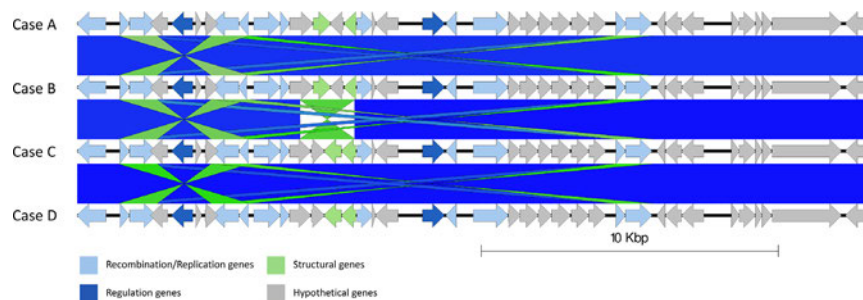


Fig. 4. Easyfig alignment of prophage 2 in all cases order descending A–D, detailing a small inversion from samples A and B relative to samples C and D (light blue). Arrow indicates gene direction. Recombination/replication genes shown in light blue, regulation associated genes in dark blue. Structure and lysis associated genes shown in light and dark green respectively, finally grey are hypothetical genes.

Within-outbreak comparison of the prophage regions of the four outbreak isolates

The chromosomes of the outbreak isolates were aligned. Genome rearrangements were identified within prophage 2, where cases A and B differ from C and D (Fig. 4), and in prophages 8 in the sequence linked to case D with respect to the other outbreak sequences (Fig. 5). In prophage 2, a 1739 bp inversion was identified involving two prophage tail genes surrounding a hypothetical gene (*yfdK*) (Fig. 4). In prophage 8, a deletion event was observed (Fig. 5). Prophage 8 was a large compound prophage (53 kpb in the sequence linked to case D and 145 kbp in the remaining samples) containing at least three separate prophages positioned sequentially without any chromosomal sequence separating them. In the sequenced linked to case D, there appears to be 92 kbp deletion (relative to the three other samples). The 92 kbp deletion contained almost two full prophage sequences, making up two sets of structural bacteriophage genes, regulatory genes, lysis genes and site-to-site recombination genes.

Within the four outbreak strains, the prophage content was equivalent except for the deleted prophage sequence in prophage 8 and the acquisition of another prophage in case D and a recombination event in prophage 2 in cases A and B relative to C and D were identified. Without a better understanding of the expected variation within prophage in STEC O157:H7 in the source population, specifically in this case the bovine gastrointestinal tract, it is difficult to be certain if these microevolutionary events represent meaningful differences between isolates. Once colonized, cattle may shed STEC O157:H7 for many months [36], and the genetic changes including the horizontal exchange of genetic information and genomic recombination/rearrangements will occur in the bacterial genomes over that time [10, 25]. Although currently little is known about the selection pressures and population dynamics of STEC O157:H7 in the bovine reservoir, microevolutionary events such as these, are unlikely to reflect a different source.

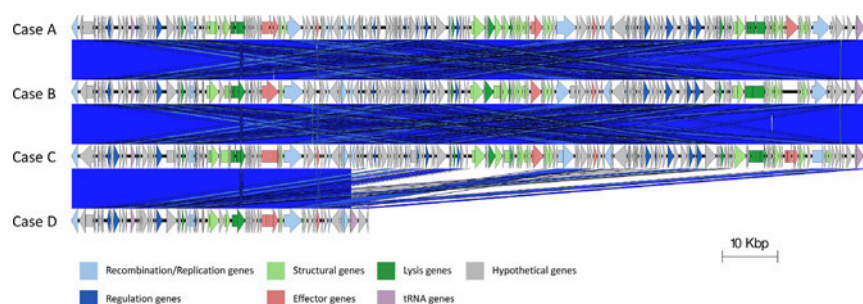


Fig. 5. Easyfig alignment of prophage 8 in all cases order descending A–D. Showing a substitution in sample D. Arrow indicates gene direction. Recombination/replication genes shown in light blue, regulation associated genes in dark blue. Effector genes are shown in pink, structure and lysis associated genes shown in light and dark green respectively and tRNAs shown as purple lines, finally grey are hypothetical genes.

SUMMARY

The advantages of using short-read WGS technologies for routine surveillance and the detection and risk management of outbreaks of STEC O157:H7 is well-established [36–38]. For example, it is unlikely that this small, nationally dispersed cluster of the cases of PT21/28, a commonly reported PT in the UK, would have been investigated prior to the implementation of WGS. However, due to the high prophage content in STEC O157:H7, assembling the genome into one contig is challenging and the utility of accessing information from the STEC accessory genome during an outbreak investigation is yet to be fully explored. In this study, we describe our methodological approach to the comparison of the accessory genomes of four temporally related cluster isolates of STEC O157:H7 epidemiologically linked to exposure to raw tripe. Comparison of Illumina with ONT sequencing data highlighted the limitations of SNP detection associated with both technologies, however, the analysis of the ONT data confirmed the close genetic relatedness demonstrated by the Illumina data. Although the within-outbreak prophage content was stable, minor structural alterations were observed in two prophages in one of the isolates. The ability to characterize the accessory genome in this way is the first step to understanding the significance of these microevolutionary events and their impact on relatedness [33, 39], the evolutionary history, virulence, and potentially the likely source and transmission [40] of this zoonotic, foodborne pathogen.

Funding information

The research was part funded by the National Institute for Health Research Health Protection Research Unit in Gastrointestinal Infections at University of Liverpool in partnership with Public Health England (PHE), in collaboration with University of East Anglia, University of Oxford and the Quadram Institute. Claire Jenkins, David Greig and Timothy Dallman are based at Public Health England. The views expressed are those of the authors and not necessarily those of the National Health Service, the NIHR, the Department of Health or Public Health England.

Acknowledgements

We would like to thank Professor David Gally at the Roslin Institute, University of Edinburgh, for his critical review of the early stages of this project.

Author contributions

T.J.D. and C. J. conceptualized the project. D.R.G. performed DNA extractions, library preparations, sequencing of isolates, data processing, genome assembly, genome polishing, genome annotation and created the Easyfig diagrams. T.J.D. performed prophage comparison using Mash and wrote associated scripts. D.R.G. performed the SNP comparison using SnapperDB and made the phylogenetic tree. D.R.G., T.J.D. and C.J., wrote the original manuscript. D.R.G., T.J.D., C.J. and S.E.G. reviewed and edited the manuscript. T.J.D. and C.J. supervised D.R.G.

Conflicts of interest

The authors declare that there are no conflicts of interest.

References

- Launders N, Byrne L, Jenkins C, Harker K, Charlett A et al. Disease severity of Shiga toxin-producing *E. coli* O157 and factors influencing the development of typical haemolytic uraemic syndrome: a retrospective cohort study, 2009–2012. *BMJ Open* 2016;6:e009933.

- Eppinger M, Mammel MK, Leclerc JE, Ravel J, Cebula TA. Genomic anatomy of *Escherichia coli* O157:H7 outbreaks. *Proc Natl Acad Sci U S A* 2011;108:20142–20147.
- Ogura Y, Mondal SI, Islam MR, Mako T, Arisawa K et al. The Shiga toxin 2 production level in enterohemorrhagic *Escherichia coli* O157:H7 is correlated with the subtypes of toxin-encoding phage. *Sci Rep* 2015;5:16663.
- Byrne L, Adams N, Jenkins C. Association between Shiga Toxin-Producing *Escherichia coli* O157:H7 *stx* gene subtype and disease severity, England, 2009–2019. *Emerg Infect Dis* 2020;26:2394–2400.
- Latif H, Li HJ, Charusanti P, Palsson Bernhard Ø, Aziz RK. A Gapless, Unambiguous genome sequence of the enterohemorrhagic *Escherichia coli* O157:H7 Strain EDL933. *Genome Announc* 2014;2:pii: e00821–14.
- Asadulghani M, Ogura Y, Ooka T, Itoh T, Sawaguchi A et al. The defective prophage pool of *Escherichia coli* O157: prophage-prophage interactions potentiate horizontal transfer of virulence determinants. *PLoS Pathog* 2009;5:e1000408.
- 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.
- Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics* 2014;30:2114–2120.
- Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* 2010;26:589–595.
- Hayashi T, Makino K, Ohnishi M, Kurokawa K, Ishii K et al. Complete genome sequence of enterohemorrhagic *Escherichia coli* O157:H7 and genomic comparison with a laboratory strain K-12. *DNA Res* 2001;8:11–22.
- McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K et al. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 2010;20:1297–1303.
- Dallman T, Ashton P, Schafer U, Jironkin A, Painset A et al. SnapperDB: a database solution for routine sequencing analysis of bacterial isolates. *Bioinformatics* 2018;34:3028–3029.
- Wick RR, Judd LM, Holt KE. Deepbiner: demultiplexing barcoded Oxford nanopore reads with deep convolutional neural networks. *PLoS Comput Biol* 2018;14:e1006583.
- De Coster W, D'Hert S, Schultz DT, Cruts M, Van Broeckhoven C. NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics* 2018;34:2666–2669.
- Wick RR. Porechop. <https://github.com/rrwick/Porechop>
- Wick RR. Filtlong. <https://github.com/rrwick/Filtlong>
- Kolmogorov M, Yuan J, Lin Y, Pevzner PA. Assembly of long, error-prone reads using repeat graphs. *Nat Biotechnol* 2019;37:540–546.
- Loman NJ, Quick J, Simpson JT. A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nat Methods* 2015;12:733–735.
- Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* 2014;9:e112963.
- H L, Handsaker B, Wysoker A, Fennell T, Ruan J et al. 1000 genome project data processing subgroup. The sequence alignment/map (SAM) format and SAMtools. *Bioinformatics* 2009;25:2078–2079.
- Vaser R, Sović I, Nagarajan N, Šikić M. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res* 2017;27:737–746.
- 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.
- Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 2014;30:2068–2069.

24. Arndt D, Grant JR, Marcu A, Sajed T, Pon A *et al.* PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic Acids Res* 2016;44:W16–W21.
25. Shaaban S, Cowley LA, McAteer SP, Jenkins C, Dallman TJ *et al.* Evolution of a zoonotic pathogen: investigating prophage diversity in enterohaemorrhagic *Escherichia coli* O157 by long-read sequencing. *Microb Genom* 2016;2:e000096.
26. Sullivan MJ, Petty NK, Beatson SA. Easyfig: a genome comparison visualizer. *Bioinformatics* 2011;27:1009–1010.
27. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH *et al.* Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol* 2016;17:132.
28. Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 2018;34:3094–3100.
29. Greig DR, Jenkins C, Gharbia S, Dallman TJ. Comparison of single-nucleotide variants identified by illumina and Oxford nanopore technologies in the context of a potential outbreak of Shiga toxin-producing *Escherichia coli*. *Gigascience* 2019;8.
30. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 2014;30:1312–1313.
31. Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA *et al.* Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Res* 2015;43:e15.
32. Wick RR, Judd LM, Holt KE. Performance of neural network basecalling tools for Oxford nanopore sequencing. *Genome Biol* 2019;20:129.
33. Greig DR, Jenkins C, Dallman TJ. A Shiga Toxin-Encoding prophage recombination event confounds the phylogenetic relationship between two isolates of *Escherichia coli* O157:H7 from the same patient. *Front Microbiol* 2020;11:588769.
34. Makino K, Ishii K, Yasunaga T, Hattori M, Yokoyama K *et al.* Complete nucleotide sequences of 93-kb and 3.3-kb plasmids of an enterohemorrhagic *Escherichia coli* O157:H7 derived from Sakai outbreak. *DNA Res* 1998;5:1–9.
35. Byrne L, Dallman TJ, Adams N, Mikhail AFW, McCarthy N *et al.* Highly Pathogenic Clone of Shiga Toxin-Producing *Escherichia coli* O157:H7, England and Wales. *Emerg Infect Dis* 2018;24:2303–2308.
36. 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.
37. Allard MW, Stevens EL, Brown EW. All for one and one for all: the true potential of whole-genome sequencing. *Lancet Infect Dis* 2019;19:683–684.
38. Herbert LJ, Vali L, Hoyle DV, Innocent G, McKendrick IJ *et al.* *E. coli* O157 on Scottish cattle farms: evidence of local spread and persistence using repeat cross-sectional data. *BMC Vet Res* 2014;10:95.
39. Cowley LA, Dallman TJ, Fitzgerald S, Irvine N, Rooney PJ *et al.* Short-term evolution of Shiga toxin-producing *Escherichia coli* O157:H7 between two food-borne outbreaks. *Microb Genom* 2016;2:e000084.
40. Greig DR, Mikhail AFW, Dallman TJ, Jenkins C. Analysis Shiga Toxin-Encoding Bacteriophage in Shiga Toxin-Producing *Escherichia coli* O157:H7 stx2a/stx2c. *Front Microbiol* 2020;11:577658.

Five reasons to publish your next article with a Microbiology Society journal

1. The Microbiology Society is a not-for-profit organization.
2. We offer fast and rigorous peer review – average time to first decision is 4–6 weeks.
3. Our journals have a global readership with subscriptions held in research institutions around the world.
4. 80% of our authors rate our submission process as 'excellent' or 'very good'.
5. Your article will be published on an interactive journal platform with advanced metrics.

Find out more and submit your article at microbiologyresearch.org.

scientific reports



OPEN

Re-analysis of an outbreak of Shiga toxin-producing *Escherichia coli* O157:H7 associated with raw drinking milk using Nanopore sequencing

David R. Greig^{1,2,3}, Vivienne Do Nascimento¹, David L. Gally³, Saheer E. Gharbia^{1,4}, Timothy J. Dallman⁵ & Claire Jenkins^{1,2}

The aim of this study was to compare Illumina and Oxford Nanopore Technology (ONT) sequencing data to quantify genetic variation to assess within-outbreak strain relatedness and characterise microevolutionary events in the accessory genomes of a cluster of 23 genetically and epidemiologically linked isolates related to an outbreak of Shiga toxin-producing *Escherichia coli* O157:H7 caused by the consumption of raw drinking milk. There were seven discrepant variants called between the two technologies, five were false-negative or false-positive variants in the Illumina data and two were false-negative calls in ONT data. After masking horizontally acquired sequences such as prophages, analysis of both short and long-read sequences revealed the 20 isolates linked to the outbreak in 2017 had a maximum SNP distance of one SNP between each other, and a maximum of five SNPs when including three additional strains identified in 2019. Analysis of the ONT data revealed a 47 kbp deletion event in a terminal compound prophage within one sample relative to the remaining samples, and a 0.65 Mbp large chromosomal rearrangement (inversion), within one sample relative to the remaining samples. Furthermore, we detected two bacteriophages encoding the highly pathogenic Shiga toxin (Stx) subtype, Stx2a. One was typical of Stx2a-phage in this sub-lineage (1c), the other was atypical and inserted into a site usually occupied by Stx2c-encoding phage. Finally, we observed an increase in the size of the pO157 IncFIB plasmid (1.6 kbp) in isolates from 2019 compared to those from 2017, due to the duplication of insertion elements within the plasmids from the more recently isolated strains. The ability to characterize the accessory genome in this way is the first step to understanding the significance of these microevolutionary events and their impact on the genome plasticity and virulence between strains of this zoonotic, foodborne pathogen.

Keywords Foodborne outbreak, Genomic epidemiology, STEC O157:H7, WGS, Nanopore, Illumina, Prophage comparison

Shiga toxin-producing *Escherichia coli* (STEC) O157:H7 is a zoonotic, gastrointestinal pathogen that colonises the gut of healthy ruminants, particularly cattle and sheep^{1,2}. Transmission to humans occurs through consumption of contaminated food or water, direct or indirect contact with animals or their environment. STEC O157:H7 infection in humans can produce a wide range of symptoms, from mild diarrhoea to severe bloody diarrhoea, abdominal cramps and vomiting³. In 5–15% of cases, the infection can lead to the development of haemolytic uremic syndrome (HUS), a severe multi-system syndrome, that can be fatal, particularly in young children and the elderly^{4,5}.

¹National Infection Service, United Kingdom Health Security Agency, London NW9 5EQ, UK. ²NIRH Health Protection Research Unit for Gastrointestinal Pathogens, Liverpool, UK. ³Division of Infection and Immunity, The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Easter Bush, Edinburgh, UK. ⁴NIHR Health Protection Research Unit in Genomes and Enabling Data, Warwick, UK. ⁵Institute for Risk Assessment Sciences (IRAS), Faculty of Veterinary Medicine, Utrecht University, 3584 CL Utrecht, The Netherlands. ✉email: david.greig@ukhsa.gov.uk

STEC O157:H7 characteristically harbours one or more bacteriophage encoded Shiga toxin genes (*stx*), belonging to one of three toxin subtypes, *stx1a*, *stx2a* and/or *stx2c*. Subtype *stx2a* is associated with causing HUS^{6,7}. STEC O157:H7 is divided into three lineages (I, II and I/II) and seven sub-lineages (Ia-Ic, IIa-IIc and I/II). Previously studies have shown that STEC O157:H7 belonging to sub-lineage Ic harbouring *stx2a* with or without *stx2c* (for the most part corresponding with phage type (PT) 21/28) was the dominant type in the UK between 1995 and 2015 and remains a common cause of STEC-HUS in England^{3,6,8,9}.

Public health surveillance of STEC O157:H7 in England, including outbreak detection and investigation, is co-ordinated by UK Health Security Agency (UKHSA), formerly Public Health England (PHE). The National Enhanced STEC Surveillance System (NESSS) integrates short-read whole genome sequencing of STEC isolates from patients with symptoms of gastrointestinal disease, with epidemiological data capturing their food and travel histories, contact with animals and other environmental exposures. The use of short-read WGS data during outbreak investigations delivers an unprecedented level of strain discrimination, facilitates case ascertainment even when epidemiological links are obscured, and provides insight into the evolutionary context for emerging pathogenic strains^{8–11}. Genetic relatedness of the sequences of the isolates is determined by reference-based variant calling to determine high-quality single nucleotide polymorphisms (SNPs)¹². Previous studies have shown that isolates with genome sequences that fall within the same 5 SNP single linkage cluster (where all samples in the cluster are linked by 5 SNPs or less) are likely to be associated with the same source^{9,11,13}.

The genome of STEC O157:H7 is approximately 5.5 Mbp in size¹⁴ and has a relatively small core genome size across the population due to a large and diverse accessory genome including the presence of large lambdoid prophages which makes up 10–15% of the STEC chromosome^{15,16}. Due to the limitations of short read sequencing technologies in handling the homologous prophage content of the STEC O157:H7 chromosome, information and context regarding inter and intra variation in prophages, structural variation and even context surrounding plasmid content is lost. With the recent development of single-molecule real time sequencing (SMRT) technologies we now have the tools to de novo assemble pathogens into individual contigs containing a single replicon¹⁷ enabling us to characterise and scrutinise the accessory genome, including prophage sequences of STEC^{17–21}.

We re-analysed the human, food and animal isolates linked to an outbreak of STEC serotype O157:H7 that occurred in 2017 on the Isle of Wight, caused by the consumption of raw drinking milk (RDM)²², using Nanopore sequencing data. In 2019, two additional cases infected with STEC O157:H7 that fell within the same 5-SNP single linkage cluster were detected¹¹. These cases were resident in the same geographical location as the farm implicated in the outbreak investigation in 2017. The aim of the study was to evaluate our methodological approach to the analysis of isolates sequenced using the Oxford Nanopore Technology (ONT) platform and to assess the accessory genome variation between isolates within an outbreak attributed to a point source exposure, and between the isolates from the cases in 2019 that were geographically related but temporally distinct.

Results and discussion

Comparison of variant calling methods

Variant calling of the Illumina sequencing data identified 23 isolates that fell within the same 5-SNP single linkage cluster. The 20 isolates linked to the outbreak in 2017 had a maximum SNP distance of 1 SNP between each other, and a maximum of 5 SNPs when including cases identified in 2019 (Fig. 1). Analysis of the Nanopore data identified a maximum of 1 SNP variants between the 20 outbreak isolates and 5 SNPs between all 23 isolates (Fig. 1).

The variants identified in the Illumina sequencing data were compared to those identified in the Nanopore data. There were seven discrepant variants called between the two technologies (Table 1). Four of the seven mismatches were determined to be false positive variants in the Illumina dataset due to ambiguously aligning short reads (Table 1). A fifth, this time false negative, variant was called in the Illumina short read data in a single sample due to low coverage over a region of the genome where the variant was called (Table 1). The remaining two discrepant variants were deemed to be false negatives in the Nanopore data due to errors associated with calling SNPs within homopolymer sequences (Table 1).

When accounting for the above false positive/false negative discrepant variants there remained only a single discrepant variant which was classified as a true variant [Nanopore (G var/T reference), Illumina (N/T reference)], accounting for the single SNP difference in sample 432,299 (Fig. 1).

This comparison highlighted the systemic differences associated with each technology, specifically the base-calling errors related to homopolymer detection observed in Nanopore data^{23,24} and the ambiguous alignment of Illumina data to homologous and paralogous regions²⁵. This comparison also demonstrates the importance of masking these regions within the reference genome to produce accurate and meaningful results²³.

Genomic features of long-read assemblies of the chromosome

The assemblies of the sequences of the 23 isolates in this study all contained either two or three closed contigs, each supporting a single replicon (Table 2). The chromosome size of the isolates sequenced in this study ($n = 23$) ranged from 5,507,151 to 5,555,878 bp in length, a maximum difference of 48,727 bp and an average size of 5,553,177 bp (Table 2).

Aligning and comparing the chromosomes of all 23 isolates within the same 5-SNP single linkage cluster led to the discovery of a large-scale recombination (LCR) event present in one of the cattle isolates (432,301) (Fig. 2). The large-scale recombination event was characterised by a 650kbp inversion between prophages 5 (*potC*) and 7 (*yebW*) (Fig. 2). At either edge of the inversion, prophages 5 (*potC*) and 7 (*yebW*) both have a 10.05kbp homologous sequence containing prophage structure encoding genes such as tail proteins, host specificity proteins and several hypothetical genes. The 10.05kbp homologous sequences in both prophages share a 97.7% sequence similarity.

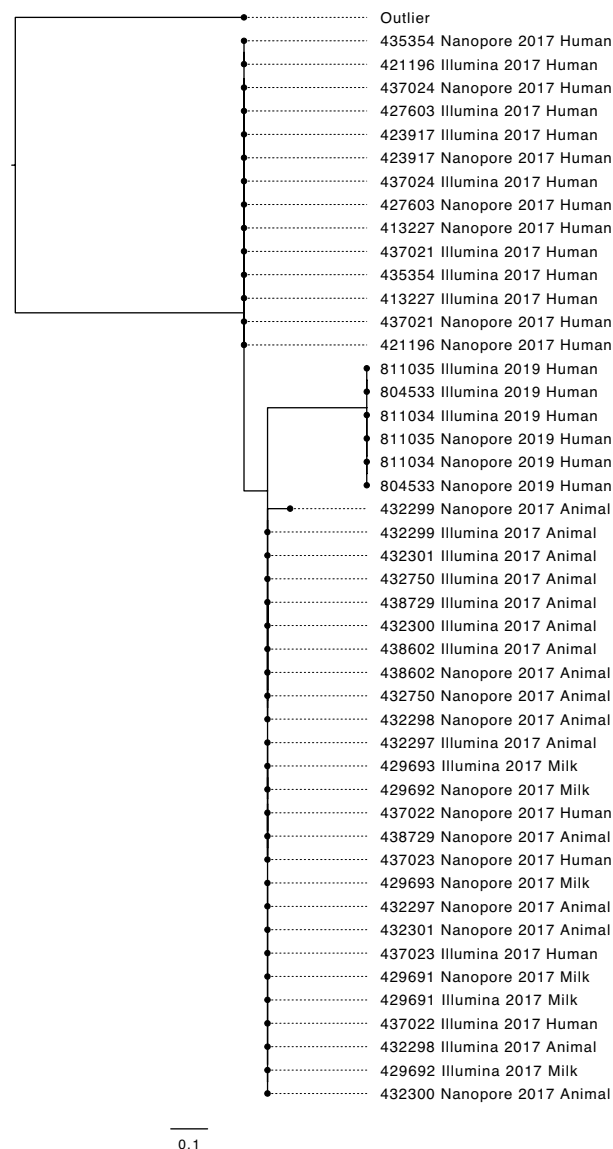


Figure 1. A maximum-likelihood phylogeny showing both Illumina derived and nanopore derived SNP-typing results for samples sequenced in this study.

It is known that the STEC O157:H7 genome undergoes large scale recombination to produce large inverted sequences within the chromosome^{26,27}. The outbreak in this study was selected because it included eight isolates from cattle, and we wanted to look for LCRs in STEC O157:H7 in the animal reservoir, as well as in the RDM and human cases. We observed a LCR in just one cattle isolate; it is uncertain whether this genetic event occurred in vivo or on sub-culture in the laboratory. Potential phenotypic effects of LCRs, for example strain fitness, infectivity, or impact on patient outcomes, are yet to be determined. Within this 5-SNP cluster of 23 isolates, minimal

Reference position	Reference base	Variant called	CDS	Locus Tag	Encodes	Count of samples with this discrepant variant (out of 23)	Reason for discrepancy (False positive/False negative)
232,300	G	A	NON CODING		n/a	1	Low coverage at position from Illumina data (FN)
270,579	A	G	R1168G	ECs0237	RhsG core protein with extension	2	Ambiguous alignment at this position from Illumina data (FP)
270,595	C	A	T1173N	ECs0237	RhsG core protein with extension	20	Ambiguous alignment at this position from Illumina data (FP)
379,516	A	G	NON CODING		n/a	4	Variant missed in homopolymer sequence by Nanopore data (FN)
418,442	T	G	E65D	ECs0395	thiogalactoside acetyltransferase	5	Variant missed in homopolymer sequence by Nanopore data (FN)
1,681,338	C	G	R314G	ECs1685	alanine racemase 2	20	Ambiguous alignment at this position from Illumina data (FP)
1,681,339	G	C	R314P	ECs1685	alanine racemase 2	20	Ambiguous alignment at this position from Illumina data (FP)

Table 1. The discrepant variant calls between illumina and nanopore datasets for all outbreak samples.

Strain ID	Source	# of contigs	Chromosome size (bp)	Total genome size (bp)	Plasmid 1 size (bp) and Inc group	Plasmid 2 size (bp) and Inc group	# of prophages	GenBank accession for chromosome	GenBank accessions for plasmids
413,227	Human	2	5,555,475	5,649,172	93,697 IncFIB	–	17	CP088060	CP088061
421,196	Human	2	5,555,066	5,649,123	94,057 IncFIB	–	17	CP088058	CP088059
423,917	Human	2	5,555,034	5,648,723	93,689 IncFIB	–	17	CP088056	CP088057
427,603	Human ^a	2	5,555,537	5,649,596	94,059 IncFIB	–	17	CP088071	CP088072
429,691	Milk	2	5,555,094	5,649,152	94,058 IncFIB	–	17	CP088054	CP088055
429,692	Milk	2	5,558,855	5,652,943	94,088 IncFIB	–	17	CP088052	CP088053
429,693	Milk	2	5,555,560	5,649,616	94,056 IncFIB	–	17	CP088050	CP088051
432,297	Animal	2	5,554,755	5,648,807	94,052 IncFIB	–	17	CP088048	CP088049
432,298	Animal	2	5,554,878	5,648,605	93,727 IncFIB	–	17	CP088046	CP088047
432,299	Animal	2	5,507,151	5,601,209	94,058 IncFIB	–	17	CP088044	CP088045
432,300	Animal	2	5,555,280	5,649,335	94,055 IncFIB	–	17	CP088042	CP088043
432,301	Animal	2	5,554,492	5,648,544	94,052 IncFIB	–	17	CP088040	CP088041
438,729	Animal	2	5,555,235	5,649,147	93,912 IncFIB	–	17	CP088069	CP088070
432,750	Animal	2	5,555,568	5,649,625	94,057 IncFIB	–	17	CP088067	CP088068
438,602	Animal	2	5,555,032	5,649,087	94,055 IncFIB	–	17	CP088038	CP088039
435,354	Human	3	5,555,067	5,735,080	94,051 IncFIB	85,962 (Inc11-γ)	17	CP088064	CP088065 + CP088066
437,021	Human ^a	2	5,555,314	5,649,370	94,056 IncFIB	–	17	CP088062	CP088063
437,022	Human ^b	2	5,554,737	5,648,789	94,052 IncFIB	–	17	CP088036	CP088037
437,023	Human ^b	2	5,554,791	5,648,842	94,051 IncFIB	–	17	CP088034	CP088035
437,024	Human ^a	2	5,555,270	5,649,330	94,060 IncFIB	–	17	CP088032	CP088033
804,533	Human	2	5,554,866	5,650,235	95,369 IncFIB	–	17	CP088030	CP088031
811,034	Human ^c	2	5,555,205	5,650,573	95,368 IncFIB	–	17	CP088028	CP088029
811,035	Human ^c	2	5,554,812	5,650,179	95,367 IncFIB	–	17	CP088026	CP088027

Table 2. Table detailing the ID of each strain, finalised assembly metrics, plasmid replicon typing, prophage counts and assembly-based accessions. ^a, ^b and ^cdetail strains shared by the same case/patient.

large-scale chromosomal variation was observed, regardless of the source (animal, food or human), the clinical outcome of the case or the year the cases were detected.

Analysis of prophage and prophage-like content

All 23 isolates contained the same number of prophages ($n = 17$) of which two were *stx2a*-encoding prophage (Figs. 2, 3). All prophages in the samples sequenced ranged from 8.2 to 144.5 kbp in size (Table S1). Across all 23

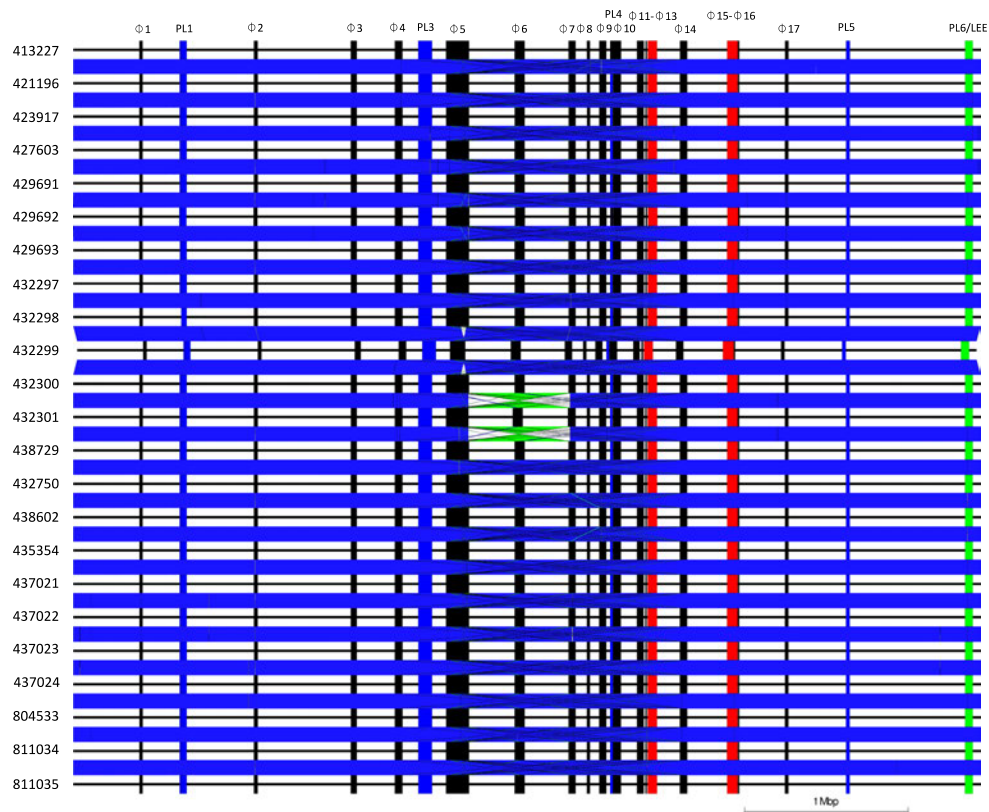


Figure 2. Easyfig alignment showing the chromosome and loci of prophages in all samples sequenced in this study. *Stx*-encoding prophage, Red; Prophage-like region, Blue; Locus of Enterocyte Effacement (LEE), Green and other non-*stx*-encoding prophages, Black.

samples, 15/17 prophages were considered the same in all samples including prophages 1 (*lexA*), 2 (tRNA-Thr), 3 (*ybhC*), 4 (*yccA*), 6 (*rspA*), 8 (*yecA*), 9 (tRNA-Ser), 10 (*ompW*), 11 (*icd*), 12 (*roxA*), 13 (*sbCB* [*stx2a*]), 14 (*yehV*), 15 (*argW* [*stx2a*]), 16 (*argW*) and 17 (*alpA*) (Figs. 2, 4).

There were 2/17 prophages that showed variation, the first type of variation was a single deletion of a 47,389bp region in a large compound prophage 5 (*potC*) in a cattle isolate (432,299), different to the isolate exhibiting the LCR (Figs. 2, 4, 5). The deletion event was surrounded by two 5.8 kbp homologous sequences containing structural tail proteins and hypothetical genes. The second type of variation observed in prophages 5 (*potC*) and 7 (*yebW*) were related to the 0.65 Mbp LCR observed where these prophages acted as the break points.

Comparing the prophages identified in the outbreak strains to those identified in a well characterised STEC O157:H7 reference genome, strain Sakai (BA000007) isolated in Japan 30 years ago, it was noted that seven prophages matched between all outbreak genomes and the reference genome and they all shared the same integration sites. Those prophages included prophage 3 and Sp3 (*ybhC*), prophage 4 and Sp4 (*yccA*), prophage 11 and Sp8 (*icd*), prophage 8 and Sp13 (*yecA*), prophage 9 and Sp14 (tRNA-Ser), prophage 16 and Sp16 (*argW*), prophage 17 and Sp17 (*alpA/ssrA*) (Figs. 2, 4A).

There were three sets of prophages that had similar regions, including prophage 6 (*rspA*) to Sp9, prophage 10 (*ompW*) to Sp12 and prophage 14 (*yehV*) to Sp15 (*stx1a*) (Figs. 2, 4A). There were a further five prophages that were unique to all 23 samples described in this study, including prophage 1 (*lexA*), prophage 2 (tRNA-Thr), prophage 12 (*roxA*), prophage 13 (*sbCB* [*stx2a*]) and prophage 15 (*argW* [*stx2a*]) (Figs. 2, 4). A further two prophages showed variation within the 23 samples sequenced in this study but did not match any prophages found in strain Sakai, including prophages 5 (*potC*) and 7 (*yebW*) (Figs. 2, 4A). Finally, Sakai prophages Sp1, Sp2, Sp5 (*stx2a*), Sp6, Sp7, Sp10, Sp11 and Sp18 were all unique to the reference strain (Fig. 4A).

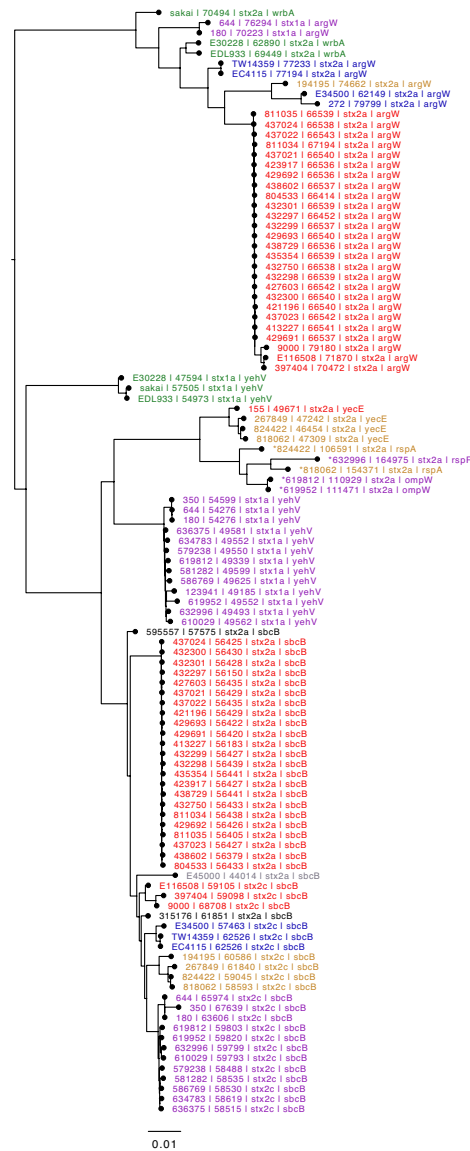


Figure 3. Neighbour joining tree based on Jaccard distances of *stx*-encoding prophages of publicly available samples and the outbreak samples sequenced in this study. Prophages are coloured by sub-lineage of STEC O157:H7. Sub-lineage Ia, Green; Ib, Yellow; Ic, Red; I/Ia, Blue; I/Ib, Grey; IIa, Orange; IIb, Black and IIc, Purple.

Comparing the prophages identified in the outbreak strains to those identified in another well characterised STEC O157:H7 UK reference genome, strain 9000 (CP018252 - 2002) which is from the same sub-lineage (Ic)

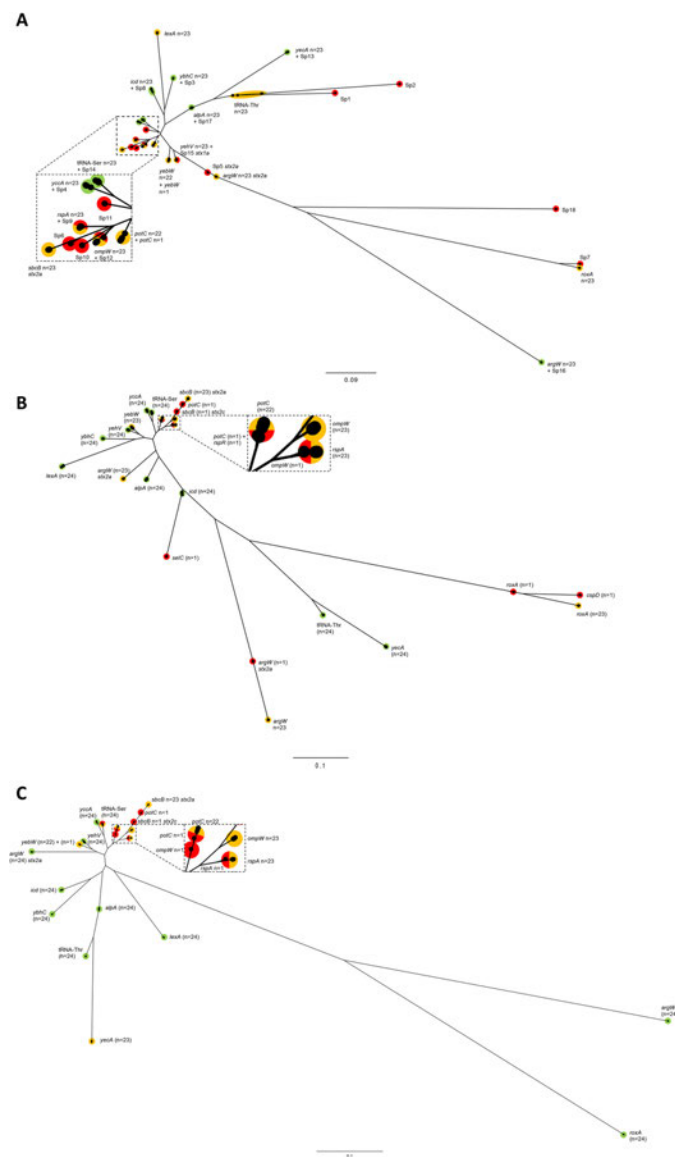


Figure 4. Mid-rooted neighbour-joining trees of Jaccard distances showing prophages from samples sequenced in this study with prophages from BA000007 (Sakai) (A), STEC O157:H7 strain 9000 (B) and STEC O157:H7 strain 397,404 (C). In each diagram prophages grouped by green are prophages shared in samples and reference genome; red are reference genome only and yellow are prophages unique samples sequenced in this study.

and phage type (PT21/28). It was noted that nine prophages were a match between the outbreak genomes and strain 9000. Those prophages included prophage 1 (*lexA*), prophage 2 (*tRNA-Thr*), prophage 3 (*ybhC*), prophage

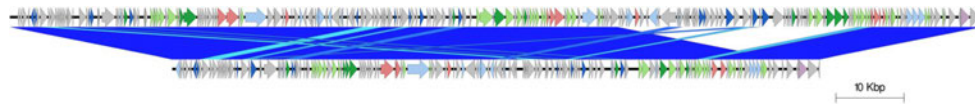


Figure 5. Easyfig alignment of compound prophage 5 with sample 432,300 (top) acting as a reference. Sample 432,299 (bottom) detailing a 47.3kbp deletion.

4 (*yccA*), prophage 8 (*yecA*), prophage 9 (tRNA-Ser), prophage 11 (*icd*), prophage 14 (*yehV*) and prophage 17 (*alpA/ssrA*) (Fig. 4B).

Finally, comparing the prophages identified in the outbreak strains compared to those identified in a recent temporally concordant outbreak in 2017 linked to handling raw pet food in the UK²⁸, from the same sub-lineage and phage type, strain 397,404 (CP043019). It was noted that ten prophages matched between all outbreak genomes and those from 397,404. Those prophages included prophage 1 (*lexA*), prophage 2 (tRNA-Thr), prophage 3 (*ybhC*), prophage 4 (*yccA*), prophage 11 (*icd*), prophage 12 (*roxA*), prophage 14 (*yehV*), prophage 15 (*argW* + *stx2a*), prophage 16 (*argW*) and prophage 17 (*alpA/ssrA*) (Fig. 4C).

Here, and in previous studies, pairwise comparisons revealed several common non-*stx*-encoding prophages found in strains of STEC O157:H7 that are temporally and geographically distinct²¹. This indicates that certain prophages are stable within the STEC O157:H7 genome and perhaps can no longer be induced. However, in contrast to the above, some prophages which have previously been described as non-inducible have been shown to be mobile^{14,16}. In the limited dataset included in our analysis to date, strains from the same sub-lineage and those more closely in time and space, had more prophages in common than those strains that were phylogenetically, geographically and temporally distinct^{21,28}. Loss and acquisition of prophage content may be influenced by both time and geographical setting.

Analysis of Shiga toxin-encoding prophages

All 23 isolates in the outbreak cluster harboured two *stx2a*-encoding prophages (Figs. 2, 3) previously undetected via Illumina sequencing; one of which was integrated at the *stx*-encoding bacteriophage integration site (SBI) *argW*, while the other was integrated at *sbcB*. The *stx2a*-encoding bacteriophage in sub-lineage Ic is commonly found at *argW*, and phylogenetic analysis showed that this bacteriophage clustered with *stx2a*-encoding phage within sub-lineage Ic (Fig. 3)²¹. The *sbcB* SBI is more commonly occupied by *stx2c*-encoding bacteriophages, and phylogenetic analysis showed that the *stx2a*-encoding bacteriophage integrated at *sbcB* in the outbreak strain was located on a branch mainly comprising *stx2c*-encoding bacteriophage²¹.

Previous studies have described the loss of *stx2c*-encoding phage and subsequent acquisition of *stx2a*-encoding phage exhibiting similar sequences to *stx2c*-encoding bacteriophage at the same SBI, in sub-lineage IIb^{10,21}, however, this is the first report of this phenomenon occurring in sub-lineage Ic. Strains of STEC O157:H7 harbouring more than one *stx2a* prophage have been described previously²⁹, but again this is the first report of the acquisition of two different *stx2a*-encoding phage in this sub-lineage. We previously showed that strains harbouring *stx2a* only belonging to sub-lineage Ic are significantly more likely to be associated with severe clinical outcomes than those strains harbouring *stx2a* only in sub-lineage IIb⁶. The representative strains of STEC O157:H7 sub-lineage IIb in previous studies¹⁰ had only one type of *stx2a*-encoding phage; the presence of two different *stx2a*-encoding phage may play a role in enhancing pathogenicity in sub-lineage Ic.

Plasmid analysis

All isolates contained an IncFIB plasmid, the pO157 that is characteristic of STEC O157:H7, ranging in size from 93,689 to 94,050 between the outbreak isolates from 2017, a maximum difference of 361bp and an average size of 93,997bp. In the three temporally distinct isolates from 2019, the size range increased to 95,369bp with a maximum difference of 1,680bp. The approximately 1.6kbp increase between IncFIB plasmids in isolates from 2019 compared to those from 2017, is due to the duplication of insertion elements within those plasmids (Fig. 6). One isolate (435,354) from 2017, also contained a smaller IncI1-γ plasmid, 85,962bp in size (Table 2a).

Conclusions

In this study, we evaluated our bioinformatics approach to analysing long read sequencing data in an outbreak setting and showed the results of these analyses correlated well with the bioinformatics pipelines routinely employed for analysing the short-read sequencing data. Minimal LCRs and/or prophage variation was observed within the isolates linked to this point source outbreak of STEC O157:H7 PT21/28 caused by the consumption

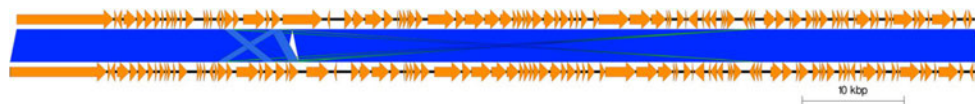


Figure 6. Easyfig alignment showing exemplar IncFIB plasmids from samples 432,301 (top) and 811,035 (bottom).

of RDM. Whether this is typical of small, geographically restricted, point source outbreaks of STEC, or characteristic of the microbiology of this specific strain and/or the epidemiological context of this setting remains to be seen. Given the association of *stx2a* with the potential to cause HUS, the discovery that the loss of *stx2c* in this strain has been followed by the acquisition of an additional *stx2a* gene at the same SBI, explains the enhanced pathogenicity associated with this clade and may represent an emerging, increased threat to public health. Supplementing routine analysis of short-read sequencing data with long-read sequencing analysis enables us to monitor the loss of, acquisition of and detection of multiple copies of *stx*-encoding bacteriophages, and improves our ability to predict emerging threats within the food chain, and provide accurate risk assessments during outbreak investigations.

Methods

Bacterial strains

There were 23 isolates of STEC O157:H7 PT21/28 in total, 20 linked to the outbreak in 2017, nine isolates were from six human cases, three from the RDM and eight from the cattle on the farm producing the milk²², and three isolates from two cases identified in 2019. All isolates belonged to sub-lineage Ic, had *stx2a* and fell within a unique five SNP single linkage cluster.

DNA extraction, library preparation, Illumina sequencing and data processing

Genomic DNA was extracted from cultures of STEC O157:H7 using the QIAasympy system (Qiagen, Hilden, Germany). The sequencing library was prepared using the Nextera XP kit (Illumina, San Diego, USA) for sequencing on the HiSeq 2500 instrument (Illumina, San Diego, USA), run with the fast protocol. FASTQ reads were processed using Trimmomatic v0.27³⁰ as previously described³¹.

DNA extraction, library preparation, Nanopore sequencing and data processing

High-molecular weight (HMW) genomic DNA was extracted and purified using the Revoluten Fire Monkey HMW DNA extraction kit (RevoluGen, UK), and DNA for each extract was quantified using a Qubit and the HS (high sensitivity) dsDNA assay kit (ThermoFisher Scientific, Waltham, USA), as previously described^{21,28}. Library preparation and sequencing was performed as previously described²⁸ before sequencing on the MinION (Oxford Nanopore Technologies, Oxford, UK) for 48 h.

Data produced in a raw FAST5 format was base-called and de-multiplexed using Guppy v3.2.10 FAST model (Oxford Nanopore Technologies, Oxford, UK) into FASTQ format. De-multiplexing was performed using Deepbiner v0.2.0³², sequencing run metrics were generated using Nanoplot v1.8.1³³, read trimming was performed using Porechop v0.2.4 (Wick RR, <https://github.com/rrwick/Porechop>)³⁴ and finally, read filtering using Filtlong v0.2.0 (Wick RR, <https://github.com/rrwick/Filtlong>)³⁵ as previously described^{21,28}.

De novo assembly, correction, re-orientation and annotation.

The filtered Nanopore FASTQ file with the 50× coverage of longest reads were assembled using Flye v2.8³⁶ with the minimum overlap length (-m) set to 10,000 and the --meta component enabled. Correction (polishing) of the assemblies was performed in a three-step process. Firstly, using Nanopolish v0.11.3³⁷, secondly, using Pilon v1.22³⁷ and finally Racon v1.3.3³⁸ as previously described^{21,28,39}. As the chromosome from each assembly was circularised and closed, they were re-orientated to start at the *dnaA* gene (GenBank accession no. NC_000913) from *E. coli* K-12, using the --fixstart parameter in Circlator v1.5.5⁴⁰. Prokka v1.13⁴¹ was used to annotate the final assemblies as previously described [21.28].

Prophage detection, excision, comparison and generation of neighbour-joining trees

Prophage sequences were detected and extracted from each samples' chromosome as described in Shaaban et al.¹⁹ and Yara et al.²¹. Prophage sequences were re-annotated using Prokka v1.13⁴¹. Mash v2.2⁴² was used to sketch (sketch length 1000, kmer length, 21) all extracted prophages in the samples sequenced in this study and all prophages found in the strain Sakai STEC O157:H7 reference genome (BA000007)¹⁴. This analysis was also performed on STEC O157:H7 PT21/28 genomes 9000 (CP018252) and 397,404 (CP043019). The pairwise Jaccard distance between the prophages was calculated and a neighbour joining tree computed for both *stx*-encoding prophages and non-*stx*-encoding prophages. Trees were visualised and annotated using FigTree v1.4.4 <https://github.com/rambaut/figtree>⁴³. Prophages and chromosomes were also aligned using Easyfig v2.2.5⁴⁴. Details for prophages from publicly available strains can be found in supplementary Table 1.

In silico plasmid typing and characterisation

The plasmid replicon for each non-chromosomal contig within each sample's final assembly was compared to PlasmidFinder's v2.1 Enterobacteriaceae reference database⁴⁵. An alignment was generated using Easyfig v2.2.5⁴⁴ relying on BLASTn v2.9⁴⁶. BLASTn parameters used were minimum identity = 90% and minimum length hit = 100bp.

Variant calling and masking, SNP typing and generation of phylogenetic trees

For reference-based variant calling both Illumina and Nanopore FASTQ reads were mapped to the Sakai STEC O157 reference genome (BA000007) using BWA v0.7.3⁴⁷ and Minimap2 v2.17⁴⁸ respectively with the use of Samtools v0.7.17⁴⁹. VCFs were produced using GATK v2.6.5 UnifiedGenotyper⁵⁰. Variants that had a high-quality SNP (> 90% for Illumina) [> 80% for Nanopore] consensus, minimum depth 10×, MQ ≥ 30) in at least one isolate were extracted for further analysis. Any variants called at positions that were within the known prophages in

Sakai were masked from further analyses. 5-methylcytosine positions were identified using Nanopolish V0.11.3¹⁷ and then masked from the Nanopore VCFs as described in Greig et al.⁵¹. The final number of positions masked was 1,189,993 bp, leaving a final reference of 4,308,457 bp.

The maximum likelihood phylogenetic tree was constructed by RAxML v8.1.17⁵² using an alignment generated from SnapperDB¹² that recombination had been accounting for by Gubbins v2.00⁵³. Visualisation of the phylogenetic tree was performed using FigTree v1.4.4⁴³. To detect false positive/negative SNPs called by Illumina and Nanopore reads, discrepant variant positions between Illumina and Nanopore relative to the reference genome were extracted. The aforementioned variants and those that also had a lower-than-average mapping quality were then masked in the alignment.

Detection and characterisation of chromosomal structural variation

Chromosome synteny was compared by aligning outbreak sample chromosomes using Easyfig v2.2.5⁴⁴. Once samples in one chromosome were aligned, structural differences could be determined and further characterised using Artemis v18.1.0⁵⁴.

To determine if there were multiple isoforms within each sample's reads (FASTQ). The FASTQ for an outbreak sample in one isoform was aligned to a finalised assembly with a different isoform using Minimap2 v2.17⁴⁸ and Samtools v0.7.17⁴⁹. Using Tablet v1.17.08.17⁵⁵, the alignments were visualised and the breakpoints at each isoform were identified. Once breakpoints were identified relative to each isoform, those positions were used with Samtools v0.7.17⁴⁹ view to isolate reads that must align across both ends of each prophage breakpoint. Any reads that did align across a given set of breakpoints must share the same size as it exists in the FASTQ file and not clipped, to be considered.

Data deposition

Illumina and Nanopore FASTQ files are available from National Center for Biotechnology Information (NCBI) BioProject PRJNA315192. The SRA (sequence read archive) accession numbers for both technologies are in supplementary Table 2. The outbreak sample finalised assemblies can also be found under BioProject PRJNA315192 and the GenBank accession numbers are located in both Table 2 and supplementary Table 2.

Data availability

All FASTQ files and assemblies were submitted to the National Centre for Biotechnology Information (NCBI). All data can be found under BioProject: PRJNA315192 - <https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA315192>. Strain-specific details can be found in Methods under data deposition.

Received: 26 September 2022; Accepted: 15 February 2024

Published online: 09 March 2024

References

- Chapman, P. A., Cerdán Malo, A. T., Ellin, M., Ashton, R. & Harkin, M. A. *Escherichia coli* O157 in cattle and sheep at slaughter, on beef and lamb carcasses and in raw beef and lamb products in South Yorkshire, UK. *Int. J. Food Microbiol.* **64**(1–2), 139–50. [https://doi.org/10.1016/S0168-1605\(00\)00453-0](https://doi.org/10.1016/S0168-1605(00)00453-0) (2001).
- Henry, M. K. et al. British *Escherichia coli* O157 in cattle study (BECS): To determine the prevalence of *E. coli* O157 in herds with cattle destined for the food chain. *Epidemiol. Infect.* **145**(15), 3168–3179. <https://doi.org/10.1017/S0950268817002151> (2017).
- Byrne, L., Jenkins, C., Launder, N., Elson, R. & Adak, G. K. The epidemiology, microbiology and clinical impact of Shiga toxin-producing *Escherichia coli* in England, 2009–2012. *Epidemiol. Infect.* **143**(16), 3475–3487. <https://doi.org/10.1017/S0950268815000746> (2015).
- Tarr, P. I., Gordon, C. A. & Chandler, W. L. Shiga-toxin-producing *Escherichia coli* and haemolytic uraemic syndrome. *The Lancet* **365**, 1073–1086. [https://doi.org/10.1016/S0140-6736\(05\)71144-2](https://doi.org/10.1016/S0140-6736(05)71144-2) (2005).
- Lynne, R. M. et al. Childhood hemolytic uraemic syndrome, United Kingdom and Ireland. *Emerg. Infect. Dis.* **11**(4), 590–596. <https://doi.org/10.3201/eid1104.040833> (2005).
- Byrne, L., Adams, N. & Jenkins, C. Association between Shiga toxin-producing *Escherichia coli* O157:H7 stx gene subtype and disease severity, England, 2009–2019. *Emerg. Infect. Dis.* **26**(10), 2394–2400. <https://doi.org/10.3201/eid2610.200319> (2020).
- EFSA Biohaz Panel et al. Pathogenicity assessment of Shiga toxin-producing *Escherichia coli* (STEC) and the public health risk posed by contamination of food with STEC. *EFSA J.* **18**(1), e05967. <https://doi.org/10.2903/j.efsa.2020.5967> (2020).
- Adams, N. L. et al. Influence of socio-economic status on Shiga toxin-producing *Escherichia coli* (STEC) infection incidence, risk factors and clinical features. *Epidemiol. Infect.* **147**, e215. <https://doi.org/10.1017/S0950268819000864> (2019).
- Dallman, T. J. et al. Whole-genome sequencing for national surveillance of Shiga toxin-producing *Escherichia coli* O157. *Clin. Infect. Dis.* **61**, 305–312. <https://doi.org/10.1093/cid/civ318> (2015).
- Byrne, L. et al. Highly pathogenic clone of Shiga toxin-producing *Escherichia coli* O157:H7, England and Wales. *Emerg. Infect. Dis.* **24**, 2303–2308. <https://doi.org/10.3201/eid2412.180409> (2018).
- Dallman, T. J., Greig, D. R., Gharbia, S. E. & Jenkins, C. Phylogenetic structure of Shiga toxin-producing *Escherichia coli* O157:H7 from sub-lineage to SNPs. *Microb. Genom.* <https://doi.org/10.1099/mgen.0.000544> (2021).
- Dallman, T. J. et al. SnapperDB: A database solution for routine sequencing analysis of bacterial isolates. *Bioinformatics* **34**, 3028–3029. <https://doi.org/10.1093/bioinformatics/bty212> (2018).
- Jenkins, C., Dallman, T. J. & Grant, K. A. 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. *Eurosurveillance* **24**(4), 1800346. <https://doi.org/10.2807/1560-7917.ES.2019.24.4.1800346> (2019).
- Hayashi, T. et al. Complete genome sequence of enterohemorrhagic *Escherichia coli* O157:H7 and genomic comparison with a laboratory strain K-12. *DNA Res.* **8**(1), 11–22. <https://doi.org/10.1093/dnares/8.1.11> (2001).
- Eppinger, M., Mammel, M. K., Leclerc, J. E., Ravel, J. & Cebula, T. A. Genomic anatomy of *Escherichia coli* O157:H7 outbreaks. *Proc. Natl. Acad. Sci. U. S. A.* **108**(50), 20142–20147. <https://doi.org/10.1073/pnas.1107176108> (2021).
- Asadulghani, M. D. et al. The defective prophage pool of *Escherichia coli* O157: Prophage-prophage interactions potentiate horizontal transfer of virulence determinants. *PLoS Pathog.* **5**(5), e1000408. <https://doi.org/10.1371/journal.ppat.1000408> (2009).

17. Loman, N. J., Quick, J. & Simpson, J. T. A complete bacterial genome assembled *de novo* using only nanopore sequencing data. *Nat. Methods* **12**, 733–735. <https://doi.org/10.1038/nmeth.3444> (2015).
18. Cowley, L. A. *et al.* Short-term evolution of Shiga toxin-producing *Escherichia coli* O157:H7 between two food-borne outbreaks. *Microb. Genom.* **2**, e000084. <https://doi.org/10.1099/mgen.0.000084> (2016).
19. Shaaban, S. *et al.* Evolution of a zoonotic pathogen: investigating prophage diversity in enterohaemorrhagic *Escherichia coli* O157 by long-read sequencing. *Microb. Gen.* **2**, e000096. <https://doi.org/10.1099/mgen.0.000096> (2016).
20. González-Escalona, N., Allard, M. A., Brown, E. W., Sharma, S. & Hoffmann, M. Nanopore sequencing for fast determination of plasmids, phages, virulence markers, and antimicrobial resistance genes in Shiga toxin-producing *Escherichia coli*. *PLoS ONE* **14**(7), e0220494. <https://doi.org/10.1371/journal.pone.0220494> (2019).
21. Yara, D. A., Greig, D. R., Gally, D. L., Dallman, T. J. & Jenkins, C. Comparison of Shiga toxin-encoding bacteriophages in highly pathogenic strains of Shiga toxin-producing *Escherichia coli* O157:H7 in the UK. *Microb. Gen.* **6**, e000334. <https://doi.org/10.1099/mgen.0.000334> (2020).
22. Treacy, J. *et al.* Outbreak of Shiga toxin-producing *Escherichia coli* O157:H7 linked to raw drinking milk resolved by rapid application of advanced pathogen characterisation methods, England, August to October 2017. *Eurosurveillance* **24**(16), 1800191. <https://doi.org/10.2807/1560-7917.ES.2019.24.16.1800191> (2019).
23. Wick, R. R., Judd, L. M. & Holt, K. E. Performance of neural network basecalling tools for Oxford Nanopore sequencing. *Genome Biol.* **20**(1), 129. <https://doi.org/10.1186/s13059-019-1727-y> (2019).
24. Dohm, J. C., Peters, P., Stralis-Pavese, N. & Himmelbauer, H. Benchmarking of long-read correction methods. *NAR Genom. Bioinform.* **2**(2), lqaa037. <https://doi.org/10.1093/nargab/lqaa037> (2020).
25. Shah, R. N. & Ruthenburg, A. J. Sequence deeper without sequencing more: Bayesian resolution of ambiguously mapped reads. *PLoS Comput. Biol.* **17**(4), e1008926. <https://doi.org/10.1371/journal.pcbi.1008926> (2021).
26. Iguchi, A., Iyoda, S., Terajima, J., Watanabe, H. & Osawa, R. Spontaneous recombination between homologous prophage regions causes large-scale inversions within the *Escherichia coli* O157:H7 chromosome. *Gene* **372**, 199–207. <https://doi.org/10.1016/j.gene.2006.01.005> (2006).
27. Fitzgerald, S. F. *et al.* Genome structural variation in *Escherichia coli* O157:H7. *Microb. Genom.* <https://doi.org/10.1099/mgen.0.000682> (2021).
28. Greig, D. R., Jenkins, C., Gharbia, S. E. & Dallman, T. J. Analysis of a small outbreak of Shiga toxin-producing *Escherichia coli* O157:H7 using long-read sequencing. *Microb. Genom.* **7**(3), mgen000545. <https://doi.org/10.1099/mgen.0.000545> (2021).
29. Greig, D. R., Mikhail, A. F. W., Dallman, T. J. & Jenkins, C. Analysis Shiga toxin-encoding bacteriophage in a rare strain of Shiga toxin-producing *Escherichia coli* O157:H7 *stx2a/stx2c*. *Front. Microbiol.* **11**, 577658. <https://doi.org/10.3389/fmicb.2020.577658> (2020).
30. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120. <https://doi.org/10.1093/bioinformatics/btu170> (2014).
31. Chattaway, M. A., Schaefer, U., Towel, R., Dallman, T. J. & Jenkins, C. Identification of *Escherichia coli* and *Shigella* Species from whole-genome sequences. *J. Clin. Microbiol.* **55**(2), 616–623. <https://doi.org/10.1128/JCM.01790-16> (2017).
32. Wick, R. R., Judd, L. M. & Holt, K. E. Deepbinner: Demultiplexing barcoded Oxford Nanopore reads with deep convolutional neural networks. *PLoS Comput. Biol.* **14**(11), e1006583. <https://doi.org/10.1371/journal.pcbi.1006583> (2018).
33. De Coster, W., D'Hert, S., Schultz, D. T., Cruts, M. & Van Broeckhoven, C. V. NanoPack: Visualizing and processing long-read sequencing data. *Bioinformatics* **34**(15), 2666–2669. <https://doi.org/10.1093/bioinformatics/bty149> (2018).
34. Wick, R. R. Porechop (2017) <https://github.com/rrwick/Porechop>.
35. Wick, R. R. Filtlong (2017) <https://github.com/rrwick/Filtlong>.
36. Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long, error-prone reads using repeat graphs. *Nat. Biotechnol.* **37**, 540–546. <https://doi.org/10.1038/s41587-019-0072-8> (2019).
37. Walker, B. J. *et al.* Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS ONE* **9**, e112963. <https://doi.org/10.1371/journal.pone.0112963> (2014).
38. Vaser, R., Sović, I., Nagarajan, N. & Šikić, M. Fast and accurate *de novo* genome assembly from long uncorrected reads. *Genome Res.* **27**, 737–746. <https://doi.org/10.1101/gr.214270.116> (2017).
39. Greig, D. R., Jenkins, C. & Dallman, T. J. A Shiga toxin-encoding prophage recombination event confounds the phylogenetic relationship between two isolates of *Escherichia coli* O157:H7 From the Same patient. *Front. Microbiol.* **11**, 588769. <https://doi.org/10.3389/fmicb.2020.588769> (2020).
40. Hunt, M. *et al.* Circlator: Automated circularization of genome assemblies using long sequencing reads. *Genome Biol.* **16**, 294. <https://doi.org/10.1186/s13059-015-0849-0> (2015).
41. Seemann, T. Prokka: Rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069. <https://doi.org/10.1093/bioinformatics/btu153> (2014).
42. Ondov, B. D. *et al.* Mash: Fast genome and metagenome distance estimation using MinHash. *Genome Biol.* **17**, 132. <https://doi.org/10.1186/s13059-016-0997-x> (2016).
43. Rambaut, A. & Drummond, A. J. FigTree (2018) <https://github.com/rambaut/figtree>
44. Sullivan, M. J., Petty, N. K. & Beatson, S. A. Easyfig: A genome comparison visualizer. *Bioinformatics* **27**, 1009–1010. <https://doi.org/10.1093/bioinformatics/btr039> (2011).
45. Carattoli, A. & Hasman, H. PlasmidFinder and In Silico pMLST: Identification and typing of plasmid replicons in whole-genome sequencing (WGS). *Methods Mol. Biol.* **2075**, 285–294. https://doi.org/10.1007/978-1-4939-9877-7_20 (2020).
46. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**(3), 403–410. [https://doi.org/10.1016/S0022-2836\(05\)80360-2](https://doi.org/10.1016/S0022-2836(05)80360-2) (1990).
47. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* **26**, 589–595. <https://doi.org/10.1093/bioinformatics/btp698> (2010).
48. Li, H. Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100. <https://doi.org/10.1093/bioinformatics/bty191> (2018).
49. Li, H. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079. <https://doi.org/10.1093/bioinformatics/btp352> (2009).
50. McKenna, A. *et al.* The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303. <https://doi.org/10.1101/gr.107524.110> (2010).
51. Greig, D. R., Jenkins, C., Gharbia, S. & Dallman, T. J. Comparison of single nucleotide variants identified by Illumina and Oxford Nanopore technologies in the context of a potential outbreak of Shiga toxin-producing *Escherichia coli*. *Gigascience* **8**(8), 23. <https://doi.org/10.1093/gigascience/giz104> (2019).
52. Stamatakis, A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313. <https://doi.org/10.1093/bioinformatics/btu033> (2014).
53. Croucher, N. J. *et al.* Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Res.* **43**, e15. <https://doi.org/10.1093/nar/gku1196> (2015).
54. Carver, T., Harris, S. R., Berriman, M., Parkhill, J. & McQuillan, J. A. Artemis: An integrated platform for visualization and analysis of high-throughput sequence-based experimental data. *Bioinformatics* **28**(4), 464–469. <https://doi.org/10.1093/bioinformatics/btr703> (2012).

55. Milne, I. *et al.* Tablet-next generation sequence assembly visualization. *Bioinformatics* 26(3), 401–402. <https://doi.org/10.1093/bioinformatics/btp666> (2010).

Author contributions

T.J.D. and C. J. conceptualised the project. V.D.N. and D.R.G. performed DNA extractions, library preparations, sequencing of isolates. D.R.G. performed data processing, genome assembly, genome polishing, genome annotation, SNP comparison using SnapperDB, prophage and plasmid comparisons, detection and characterisation of structural variation and created the figures including phylogenetic trees. T.J.D. wrote associated bioinformatics scripts. D.R.G., T.J.D. and C.J., wrote the original manuscript. D.R.G., T.J.D., D.L.G., C.J. and S.E.G. reviewed and edited the manuscript. D.L.G., T.J.D. and C.J. supervised D.R.G.

Funding

The research was part funded by the National Institute for Health Research (NIHR) Health Protection Research Unit (HPRU) in Gastrointestinal Infections at the University of Liverpool (UK), in partnership with UK Health Security Agency (UKHSA) formally Public Health England (PHE), in collaboration with the University of Warwick. C.J. and D.R.G. are based at UKHSA. The views expressed are those of the authors and not necessarily those of the National Health Service, the NIHR, the Department of Health nor UKHSA. Health Protection Research Unit (HPRU) in Genomics and Enabling Data is a collaboration funded by the National Institute for Health Research (NIHR) between Public Health England (PHE), the University of Warwick, the Centre for Genomic Pathogen Surveillance and the University of Cambridge.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-024-54662-0>.

Correspondence and requests for materials should be addressed to D.R.G.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© Crown 2024

Surveillance of antimicrobial resistant Shiga toxin-producing *E. coli* O157:H7 in England, 2016–2020

David R. Greig, Vivienne Do Nascimento, Israel Olonade, Craig Swift, Satheesh Nair and Claire Jenkins*

Gastrointestinal Bacteria Reference Unit, UK Health Security Agency, 61 Colindale Avenue, London NW9 5EQ, UK

*Corresponding author. E-mail: claire.jenkins1@ukhsa.gov.uk

Received 15 February 2023; accepted 8 July 2023

Objectives: Shiga toxin-producing *Escherichia coli* (STEC) O157:H7 are zoonotic pathogens and transmission to humans occurs via contaminated food or contact with infected animals. The aim of this study was to describe the frequency, and distribution across the phylogeny, of antimicrobial resistance (AMR) determinants in STEC O157:H7 isolated from human cases in England.

Methods: Short-read whole-genome sequencing data from 1473 isolates of STEC O157:H7 from all seven sub-lineages (Ia-Ic, IIa-IIc and I/II) were mapped to genes known to confer phenotypic resistance to 10 different classes of antibiotic. Long-read sequencing was used to determine the location and genomic architecture of the AMR determinants within phylogenetic clusters exhibiting multidrug resistance.

Results: Overall, 216/1473 (14.7%) isolates had at least one AMR determinant, although the proportion of isolates exhibiting AMR varied by sub-lineage. The highest proportion of AMR determinants were detected in sub-lineages Ib (28/64, 43.7%), I/II (18/51, 35.3%) and IIc (122/440, 27.7%). In all sub-lineages, the most commonly detected AMR determinants conferred resistance to the aminoglycosides, tetracyclines and sulphonamides, while AMR determinants conferring resistance to fluoroquinolones, macrolides and third-generation cephalosporins were rarely detected. Long-read sequencing analysis showed that the AMR determinants were co-located on the chromosome in sub-lineages Ib and lineage I/II, whereas those associated with sub-lineage IIc were encoded on the chromosome and/or large plasmids.

Conclusions: AMR genes were unevenly distributed across the different sub-lineages of STEC O157:H7 and between different clades within the same sub-lineage. Long-read sequencing facilitates tracking the transmission of AMR at the pathogen and mobile genetic element level.

Introduction

Surveillance of antimicrobial resistance (AMR) detected in gastrointestinal (GI) pathogens is important from both the clinical and public health perspective. Data on the prevalence of resistance to the different classes of antimicrobials informs guidance on the clinical management and empirical treatment of patients presenting with GI symptoms (<https://www.bmj.com/content/372/bmj.n437>, <https://bestpractice.bmj.com/topics/en-gb/1174>). Monitoring of AMR contributes to the tracking of the global spread of multidrug resistant (MDR) GI pathogens isolated from patients reporting traveller's diarrhoea and provides insight in to emerging MDR, and novel exposures and routes of transmission.^{1–6} Surveillance of zoonotic, foodborne GI pathogens such as Shiga toxin-producing *Escherichia coli* (STEC), facilitates monitoring the transmission of AMR from the animal reservoir to humans via the food chain.^{7,8} The analysis of short and long-read genome

sequencing data for routine surveillance of GI pathogens enables us to track the acquisition and intra- and inter-species dissemination of AMR determinates on mobile genetic elements (MGE).^{9–15}

STEC is known to colonize the gastrointestinal tract of healthy ruminants, and cattle studies have shown that approximately 20% of cattle herds are colonized with STEC O157:H7, the most frequently detected STEC serotype in the UK.¹⁶ Transmission to humans occurs via the faecal-oral route, either through consumption of contaminated food or through contact with animals or their environments.

There are three main lineages of STEC O157:H7 (I, II and I/II) and seven sub-lineages, Ia-Ic, IIa-IIc and I/II.^{17,18} Historically, the early outbreaks of STEC O157:H7 in the 1980s were caused by isolates belonging to lineage I/II,^{17,18} however, over the last 10 years the dominant sub-lineages have been Ic, IIa, IIb and IIc. By contrast, strains belonging to Ia, Ib and I/II have been rarely detected.^{17,19} In the UK, sub-lineages Ic and IIb are almost

© Crown copyright 2023.

This article contains public sector information licensed under the Open Government Licence v3.0 (<http://www.nationalarchives.gov.uk/doc/open-government-licence/version/3/>).

1 of 11

Downloaded from <https://academic.oup.com/jac/advance-article/doi/10.1093/jac/dkad231/7237980> by UK Health Security Agency (UKHSA) user on 14 August 2023

entirely linked to domestic acquisition, whereas sub-lineages IIa and IIc are characterized by both domestic and travel-acquired isolates.¹⁹

Prior to the implementation of WGS at UK Health Security Agency (UKHSA), monitoring AMR in STEC was limited, partly because antibiotic treatment of STEC is contraindicated and partly because of the laboratory safety implications of performing extra testing on Hazard Group 3 pathogens. Since 2015, genome derived AMR profiles of all STEC submitted to UKHSA have been available in real-time.^{7,8} The aim of this study was to describe the occurrence and frequency of AMR determinants in STEC O157:H7 isolates linked to cases resident in England and determine the distribution of AMR determinants across the STEC O157:H7 phylogeny.

Methods

Bacterial strains

In England, faecal specimens from patients with suspected gastrointestinal infection are tested for a range of gastrointestinal pathogens, including STEC O157:H7 [UK SMI S 7: gastroenteritis - GOV.UK (www.gov.uk)]. Isolates are submitted to UKHSA for WGS.^{17,19}

Whole-genome sequencing on the Illumina platform

Genomic DNA was extracted and sequenced on an Illumina HiSeq platform to produce 100 bp short-read sequence fragments (Illumina, Cambridge, UK). AMR determinants were sought using 'Gene-Finder', a customized algorithm that uses Bowtie2 (v.2.3.5.1)²⁰ to map reads to a set of reference sequences and Samtools (v.1.8)²¹ to generate an mpileup file, as previously described.^{7,8} The presence of resistance genes was defined based on 100% read coverage and >90% nucleotide identity relative to the reference sequence, with the exception of β -lactamase variants that were determined with 100% identity using the reference sequences downloaded from the Lahey (www.lahey.org) and National Center for Biotechnology Information (NCBI) β -lactamase data resources (https://www.ncbi.nlm.nih.gov/pathogens/beta-lactamase-data-resources). Known acquired-resistance genes and resistance-conferring mutations relevant to β -lactams (including carbapenems), fluoroquinolones, aminoglycosides, chloramphenicol, macrolides, sulphonamides, tetracyclines, trimethoprim, rifamycins and fosfomycin were included in the analysis. Chromosomal mutations focused on variations in the quinolone resistant determining regions (QRDR)s of *gyrA*, *gyrB*, *parC* and *parE*. Isolates that had AMR determinants known to confer resistance to three or more classes of antimicrobial were defined as MDR.

Illumina reads were mapped to the STEC O157:H7 reference genome Sakai (GenBank accession BA000007) using BWA-MEM v.0.7.13.²² SNPs were identified using GATK v.2.6 in unified genotyper mode.²³ Core-genome positions that had a high-quality SNP (>90% consensus, minimum depth 10× MQ ≥30) in at least one isolate were extracted for further analysis. Genomes were compared to the sequences held in the UKHSA STEC O157:H7 WGS database, using SnapperDB v.0.2.5.²⁴ The maximum-likelihood phylogenetic tree was constructed by RAxML v.8.1.17²⁵ using an alignment generated from SnapperDB v.0.2.5²⁴ in which recombination had been masked by Gubbins v.2.00.²⁶ Visualization/annotation of the phylogenetic tree was performed using FigTree v.1.4.4.

Nanopore-based whole-genome sequencing

Genomic DNA was extracted and purified using the Revolugen Fire Monkey DNA extraction kit (Revolugen, Glossop, UK). Library preparation was performed using the rapid barcoding kit (SQK-RBK004) (Oxford

Nanopore Technologies, Oxford, UK). The prepared libraries were loaded onto FLO-MIN106 R9.4.1 flow cells (Oxford Nanopore Technologies, Oxford, UK) and sequenced using the MinION (Oxford Nanopore Technologies, Oxford, UK) for 48 h.

Data produced in a raw FAST5 format were base-called and demultiplexed using the Guppy v.4.3.4 FAST model (Oxford Nanopore Technologies, Oxford, UK) into FASTQ format and grouped in each sample's respective barcode. Demultiplexing was performed using Deepbinner v.0.2.0.²⁷ Sequencing run metrics were generated using Nanoplot v.1.8.1.²⁸ The barcode and y-adaptor from each sample's reads were trimmed, and chimeric reads split using Porechop v.0.2.4 (Wick RR, https://github.com/rwrick/Porechop). Finally, the trimmed reads were filtered using FilTlong v.0.2.0 (Wick RR, https://github.com/rwrick/FilTlong) with the following parameters, min length=1000 bp, keep percentage=90 and target bases=275 Mbp, to generate approximately 50× coverage of the STEC genome (5.5 Mbp) to generate two FASTQ files one for the longest (–length_weight=10) and one for the highest-quality (–mean-q-weight=10) reads.

The filtered nanopore FASTQ file with the 50× coverage of longest reads was assembled using Flye v.2.8²⁹ with the minimum overlap length (–m) set to 10000 and the –meta component enabled. Assembly correction (or polishing) was performed in a three-step process. First, correction was initiated using Nanopolish v.0.11.3³⁰ using both the highest-quality nanopore FASTQ and the FAST5 files for each respective sample accounting for methylation using the –methylation-aware=dcn and –min-candidate-frequency=0.1. The alignment of reads to draft assembly was generated using Minimap2 v.2.17.³¹ Secondly, the correction was continued with Pilon v.1.22³² using Illumina FASTQ reads as the query dataset with the use of BWA v.0.7.17 and Samtools v.1.7. Finally, Racon v.1.3.3, also using BWA v.0.7.17²² and Samtools v.1.7,²¹ was used again with the Illumina FASTQ reads. As the chromosome from each assembly was circularized and closed, they were re-orientated to start at the *dnaA* gene (GenBank accession no. NC_000913) from *E. coli* K-12, using the –fixstart parameter in Circlator v.1.5.5³³ and Prokka v.1.13³⁴ with default parameters.

In silico plasmid replicon typing and detection of and characterization of AMR determinants

The plasmid replicon for each non-chromosomal contig within the final assembly of each sample was performed using PlasmidFinder v.2.1³⁵ with the Enterobacteriaceae, minimum identity=90% and minimum coverage=90% parameters set. The MGE harbouring AMR determinants where annotated using PGAP build 2022-12-13.³⁶ Finally, alignments were generated using Clinker v.0.0.27.³⁷

Data deposition

Illumina FASTQ files are available from NCBI BioProject PRJNA315192. Nanopore FASTQ and finalized assembly files are also available from BioProject PRJNA315192. Short-read archive accession numbers for each isolate are listed in the [Supplementary Table S1](#) (available as [Supplementary data](#) at JAC Online).

Results

Overall, 217/1473 (14.7%) isolates in this study had at least one AMR determinant in the Gene-Finder database (Table 1 and [Supplementary Table S1](#)). However, the proportion of isolates exhibiting full susceptibility varied by sub-lineage, with the highest proportion of AMR determinants detected in sub-lineages Ib (28/64, 43.7%), I/II (18/51, 35.3%) and IIc (122/440, 27.7%), and the lowest proportion detected in sub-lineages Ic (7/342, 2.0%), IIb (7/208, 3.3%), Ia (3/38, 7.9%) and IIa (31/323, 9.6%) (Table 1 and Figures 1–3). The most commonly detected AMR

Table 1. Proportion of isolates within each lineage and sub-lineage predicted to confer resistance to a range of classes of antimicrobials, and the proportion of cases reporting recent travel (within 7 days of onset of symptoms) outside the UK

Lineage	Ia		Ib		Ic		IIa		IIb		IIc		I/II	
	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%
Total	38		64		342		323		215		440		51	51
AMR														
Fully susceptible	35	92.1	36	56.3	335	98	292	90.4	208	96.7	318	72.3	33	64.7
B-lactams	0	0	12	18.8	5	1.5	13	4	4	1.9	95	21.5	0	0
Aminoglycosides	3	7.9	27	42.2	0	0	17	5.3	2	0.9	108	24.5	15	29.4
Fluroquinolones	1	2.6	1	1.6	0	0	7	2.2	0	0	32	7.3	2	2
Macrolides	3	7.9	0	0	0	0	1	0.3	0	0	7	1.6	0	0
Trimethoprim	1	2.6	14	21.9	0	0	6	1.9	0	0	96	21.8	0	0
Sulphonamides	2	5.3	27	42.2	5	1.5	17	5.3	0	0	107	24.3	15	29.4
Tetracycline	3	7.9	26	40.6	4	1.2	17	5.3	1	0.5	99	22.5	16	31.4
Choramphenicol	0	0	0	0	0	0	9	2.8	0	0	48	10.9	1	1
Recent travel outside UK	11	28.9	18	28.1	15	4.4	86	26.6	9	4.2	93	21.1	8	15.7

Travel history was captured from information on the laboratory submission form; if no travel history was reported, it cannot be inferred that the case did not travel.

determinants across all seven sub-lineages were to the aminoglycosides, tetracyclines and sulphonamides, while AMR to fluroquinolones, macrolides and third-generation cephalosporins was rare (Table 1 and Supplementary Figure S1).

Resistance to β -lactams

Of the 1473 isolates in this study, 129 (8.8%) had genes predicted to confer resistance to β -lactams. Including *bla*_{TEM-1} (*n*=116), *bla*_{TEM-190} (*n*=1), *bla*_{TEM-191} (*n*=3), *bla*_{TEM-117} (*n*=2) and the extended-spectrum β -lactamases (ESBLs) *bla*_{CTX-M-15} (*n*=7) (Table 2). Carbapenemase genes were not detected in any of the isolates. The highest proportion of STEC O157:H7 isolates resistant to β -lactams belonged to either sub-lineage Ib (12/64, 18.8%) or I/c (95/441, 21.5%) (Table 2 and Figures 2 and 3). The seven isolates harbouring *bla*_{CTX-M-15} were distributed across three sub-lineages, Ib (*n*=1), IIa (*n*=2) and I/c (*n*=4); three cases reported travelling outside the UK within 7 days of onset of symptoms to the Middle East, Egypt and Spain (Supplementary Table S1).

Resistance to tetracyclines, sulphonamides and trimethoprim

There were 166/1473 (11.3%) isolates dispersed across all seven sub-lineages that had *tetA*, predicted to confer tetracycline resistance. The sub-lineages with the highest proportions of isolates exhibiting predicted resistance to tetracycline were sub-lineages Ib (26/64, 40.6%), I/II (16/50, 32.0%) and I/c (99/441, 22.5%) (Tables 1 and 2).

One hundred and seventy-three (11.7%) isolates had genes expected to confer sulphonamide resistance, found in all but one of the sub-lineages. Of these, one had *sul-1*, 111 had *sul-2*, 53 had both *sul-1* and *sul-2*, and nine had *sul3* either alone (*n*=1), or in combination with *sul2* (*n*=6) or *sul1* and *sul2* (*n*=2) (Table 2). As with tetracycline, the sub-lineages with the highest proportions of isolates exhibiting predicted resistance to the sulphonamides were sub-lineages Ib (14/64, 21.8%), I/II

(15/51, 29.4%) and I/c (59/440, 13.4%) (Table 1 and Figures 1–3). All isolates harbouring *sul-3* belonged to sub-lineage I/c.

One hundred and seventeen (7.9%) isolates had *dfrA* genes (*dfrA-1*=66; *dfrA-5*=1; *dfrA-8*=41; *dfrA-12*=10, *dfrA-14*=1), conferring resistance to trimethoprim, either alone or in combination (Table 1). The highest proportions of isolates exhibiting predicted resistance to trimethoprim were sub-lineages Ib (14/64, 21.9%) and I/c (59/440, 13.4%), with all trimethoprim resistant isolates in Ib harbouring *dfrA-1*, whereas those in I/c had *dfrA-1*, *dfrA-8* and *dfrA-12* (Table 2).

Resistance to aminoglycosides

There were 172/1472 (11.7%) isolates that had genes expected to confer streptomycin resistance, of which 167/172 (70.9%) had *strA*, *strB* only or in combination with another gene known to confer resistance to one of the aminoglycosides, and four had *aadA* variants (*aadA-1*, *n*=2; *aadA-1b*, *aadA-2*) in the absence of *strA*, *strB*, and one had *aac6'-Iy* only (Table 2). Most isolates harbouring aminoglycoside resistance genes in sub-lineages Ia, Ib, Ic, IIa and I/II had *strA*, *strB* alone (62/67, 92.5%), compared to sub-lineage I/c where 65/108 (60.2%) isolates had *strA*, *strB* alone. A wide variety of genes predicted to confer resistance to gentamicin, kanamycin, tobramycin, neomycin and apramycin, were detected in sub-lineage I/c, and these are listed in Table 2.

Resistance to quinolones, macrolides and the phenicols

Forty-three STEC O157 isolates (43/1473, 2.9%), distributed across sub-lineages IIa (11/232), Ib (3/64) and I/c (32/441), contained mutations in *gyrA/parC* (predominantly *gyrA83:S-L*) and/or harboured a plasmid-mediated quinolone resistance genes. There were 25 isolates that had *gyrA83:S-L* only, four that had *gyrA83:S-L*; *parC80:S-I* or *gyrA83:S-L* plus *qnrS1* and five had *qnrS1* only (Table 2).

Resistance determinants predicted to confer to the macrolides were detected in eight isolates (8/1472). Five had *ermB*

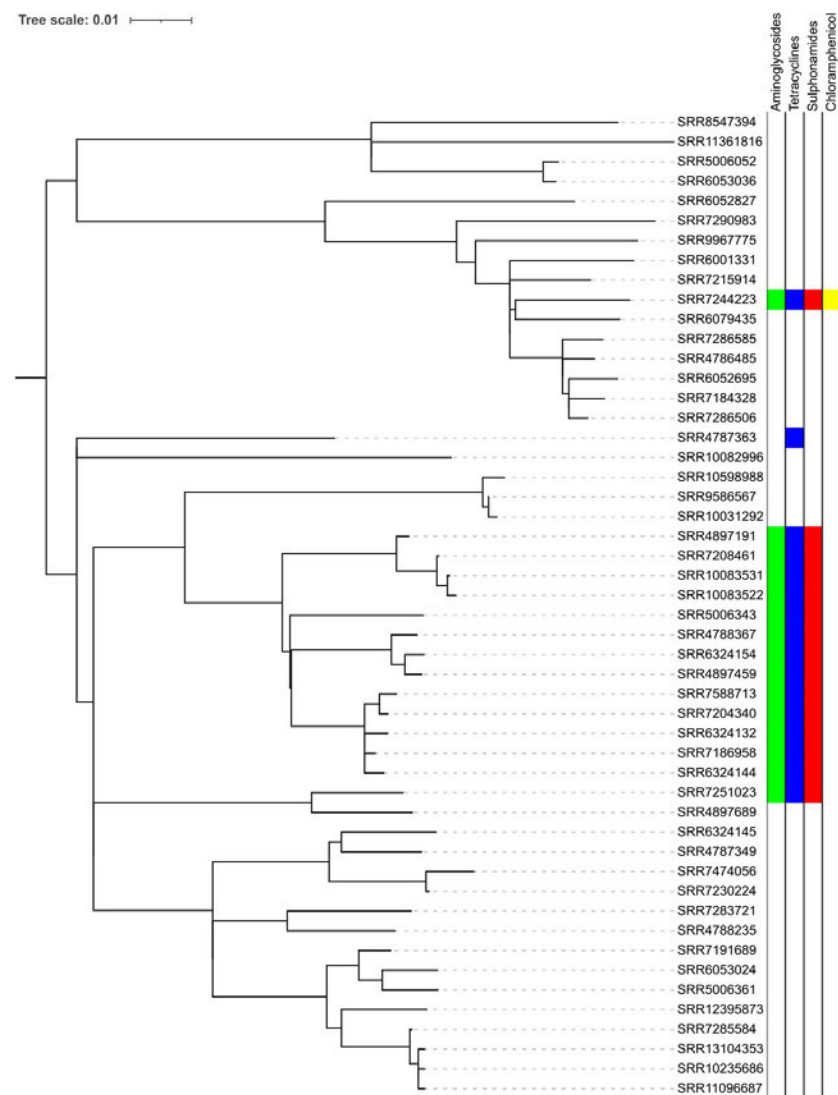


Figure 1. Phylogeny of STEC O157:H7 lineage I/II showing the distribution of AMR determinants.

and all belonged to sub-lineage IIc and three had *mphA* (sub-lineages IIa $n=2$ and IIc $n=1$) (Table 2). Fifty-one isolates had *floR*, *catA* or *cml1* either alone or in combination (*floR*, *cat-A* or *cml-1*). All but one of these isolates belonged to either sub-lineage IIa or IIc (Table 2).

Phylogenetic analysis of AMR isolates and analysis of ONT data

The sub-lineages exhibiting the highest proportion of MDR isolates (lineage I/II and sub-lineages Ib and IIc) were analysed

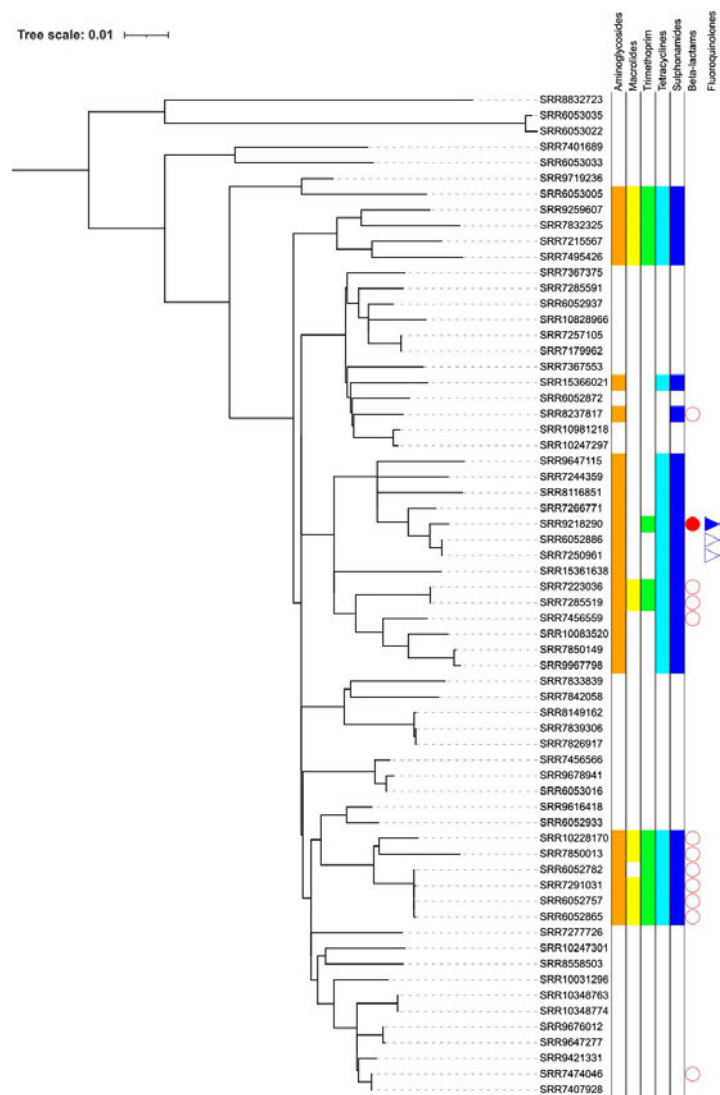


Figure 2. Phylogeny of STEC O157:H7 lineage Ib showing the distribution of AMR determinants.

further (Figure 1). Most MDR isolates in lineage I/II were restricted to a 50 SNP single linkage cluster and had *strA*, *strB/tetA/sul2* (Figure 1). Long-read sequencing revealed that these three AMR determinants were co-located on a 9kbp Tn10 intergron

integrated into the chromosome at 2080998–2089926 within a prophage at *clpA* (Figure 4). Two isolates each had a single mutation in *gyrA83:S-L*, both were from cases reporting recent travel to Mexico (Supplementary Table S1).

Downloaded from https://academic.oup.com/jac/advance-article/doi/10.1093/jac/dkac231/7237980 by UK Health Security Agency (UKHSA) user on 14 August 2023

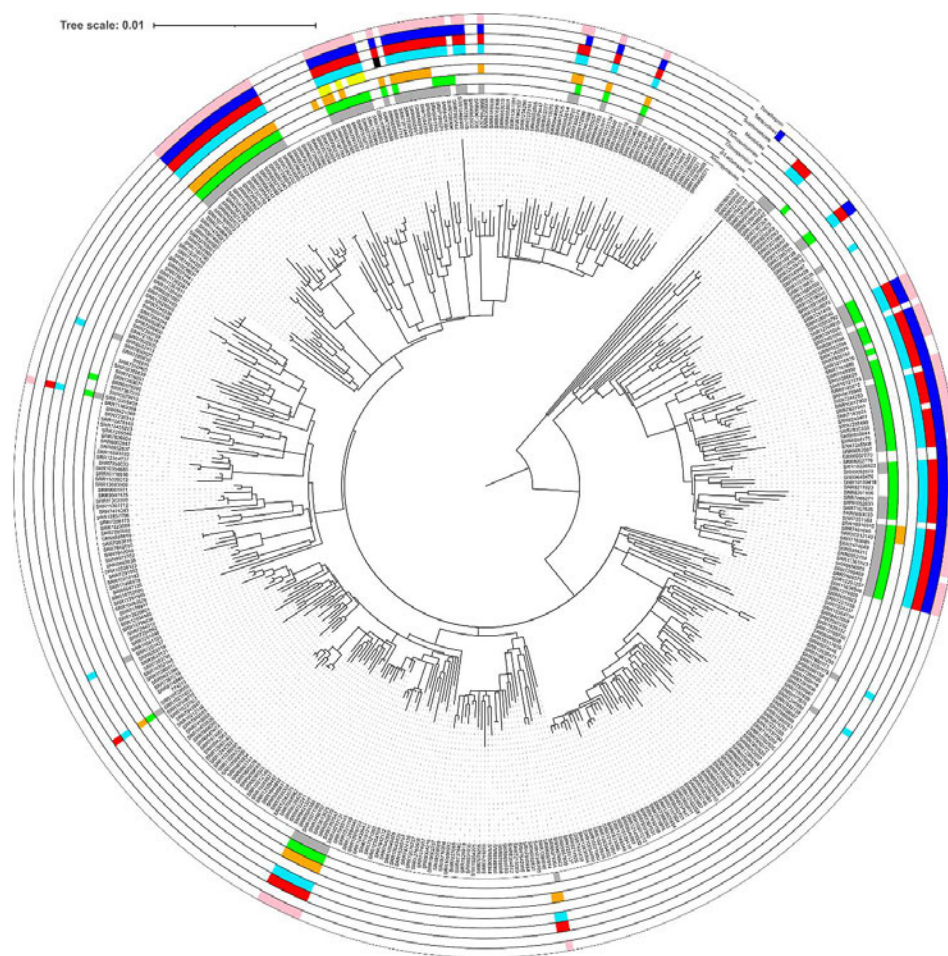


Figure 3. Phylogeny of STEC O157:H7 lineage IIc showing the distribution of AMR determinants.

Most MDR isolates that fell within sub-lineage Ib had either *strA-strB/tetA/sul2* or *bla_{TEM-1}/strA-strB/dfrA-1/tetA/sul1/sul2* (Figure 2). Acquisition of *dfrA-1* correlated with acquisition of *sul-1* and, in some isolates, *bla_{TEM-1}* (Figure 2). Again, all the AMR determinants were co-located on the chromosome on a 39 466 bp fragment of DNA inserted within the LEE at 5 376 390–5 415 856 (Figure 5).

The AMR profiles in sub-lineage IIc were the most variable. Loss and acquisition of a wide range of different AMR determinants was observed across the phylogeny, although most of the MDR isolates belonged to one of four 50 SNP single linkage

clusters and one 100 SNP single linkage cluster, each with its own characteristic AMR profile (Figure 3).

MDR isolates belonging to Cluster1 had the AMR profile *bla_{TEM-1}/strA, strB/aadA-8b/aadA-2/aac(3)-II/dfrA-12/sul-2/floR/cata-1*, predicted to confer resistance to aminoglycosides, ampicillin, trimethoprim, sulphonamides and chloramphenicol. All these AMR determinants were co-located on the chromosome on a 30 071 bp fragment of DNA that also encodes the mercury resistance cassette, inserted into the genome at position 1 555 382–1 585 453.

Table 2. Number of isolates within each lineage and sub-lineage harbouring AMR determinants included in the UKHSA Gene-Finder reference database, predicted to confer resistance to a range of classes of antimicrobials

AMR determinant	Ia n	%	Ib n	%	Ic n	%	IIa n	%	IIb n	%	IIc n	%	I/II n	%
<i>B-lactams</i>														
TEM-1	0		11	17.2	5	1.5	9	2.7	3	1.4	90	20.5	0	
CTX-M-15	0		1	1.6	0		2	0.6	0		4	0.9	0	
TEM-117	0		0		0		1	0.3	1	0.5	0		0	
TEM-190	0		0		0		1	0.3	0		0		0	
TEM-191	0		0		0		0		0		6	1.4	0	
DHA-21	0		0		0		0		0		1	0.2	0	
<i>Aminoglycosides</i>														
strA-strB	3	7.9	27	42.2	6	1.8	17	5.3	1	0.5	105	23.9	15	29.4
aac(3)-IId	0		1	1.6	0		0		0		7	1.6	0	
aadA-1b	0		0		0		2	0.6	0		5	1.1	0	
aadA-2	0		0		0		2	0.6	0		12	2.7	0	
aadA-17	0		0		0		1	0.3	0		0		0	
aac(6)-Iy	0		0		0		0		1	0.5	0		0	
aac(3)-Iva	0		0		0		0		1	0.5	23	5.2	0	
aadD	0		0		0		0		1	0.5	23	5.2	0	
aph(4)-Ia	0		0		0		0		1	0.5	22	5	0	
aadA-8b	0		0		0		0		0		9	2	0	
aadA-3	0		0		0		0		0		2	0.5	0	
aac(3)-IIa	0		0		0		0		0		2	0.5	0	
aadA-1	0		0		0		0		0		2	0.5	0	
<i>Fluoroquinolones</i>														
gyrA83:S-L	1	2.6	0		0		3	0.9	0		26	5.9	2	3.9
gyrA83:S-A	0		0		0		1	0.3	0		0		0	
qnrS-1	0		1	1.6	0		4	1.2	0		6	1.4	0	
qnrB-19	0		3	4.7	0		2	0.6	0		0		0	
gyrA83:S-L;parC80:S-I	0		0		0		0		0		4	0.9	0	
<i>Macrolides</i>														
mphA	0		0		0		1	0.3	0		2	0.5	0	
mphB	0		12	18.8	0		0		0		34	7.7	0	
ermB	0		0		0		0		0		3	0.7	0	
erm42	0		0		0		0		0		2	0.5	0	
<i>Trimethoprim</i>														
dfrA-8	1	2.6	0		0		0		0		40	9.1	0	
dfrA-1	0		14	21.9	0		3	0.9	0		49	11.2	0	
dfrA-5	0		0		0		1	0.3	0		0		0	
dfrA-14	0		0		0		3	0.9	0		0		0	
dfrA-12	0		0		0		0		0		10	2.3	0	
<i>Tetracycline</i>														
tetA	3	7.9	26	40.6	4	1.2	17	5.3	1	0.5	99	22.5	16	31.3
<i>Sulphonamide</i>														
sul1	0		13	20.3	0		0		0		48	10.9	0	
sul2	2	5.2	27	42.2	5	1.5	16	5	0		101	23	15	29.4
sul3	0		0		0		2	6.2	0		8	1.8	0	
<i>Chloramphenicol</i>														
catA	0		0		0		0		0		15	3.4	0	
floR	0		0		0		8	2.5	0		43	9.8	1	2
cml-1	0		0		0		1	0.3	0		7	1.6	0	
Fully susceptible	35	92.1	36	56.3	335	98	292	90.4	208	96.7	318	72.3	33	64.7
At least 1 AMR determinant	3	7.9	28	43.7	7	2	31	9.6	7	3.3	122	27.7	18	35.3
Total	38		64		342		323		215		440		51	

Downloaded from https://academic.oup.com/jac/advance-article/doi/10.1093/jac/dkx231/7237980 by UK Health Security Agency (UKHSA) user on 14 August 2023

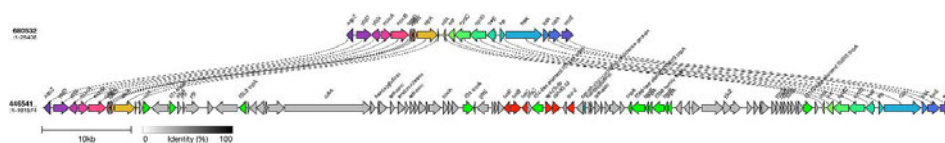


Figure 4. Long-read sequencing analysis of the region encoding the AMR determinants in lineage I/II showed that *strA*, *strB/tetA/sul2* were co-located on a 9kbp Tn10 intergron integrated into the chromosome at 2 080 998–2 089 926 within a prophage at *clpA*.

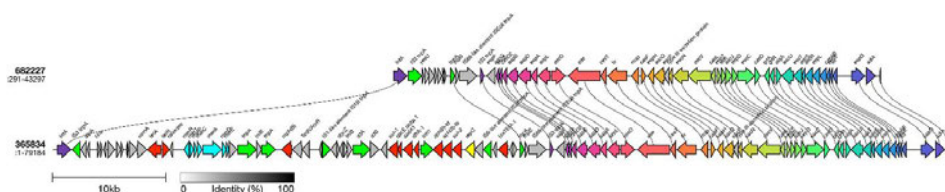


Figure 5. Long-read sequencing analysis of the region encoding the AMR determinants in sub-lineage Ib showed that all the AMR determinants (either *strA*-*strB/tetA/sul2*, or *bla*_{TEM-1}/*strA*-*strB/dfrA-1/tetA/sul1/sul2*) were co-located on the chromosome on a 39 466 bp fragment of DNA inserted within the Locus of Enterocyte Effacement at 5 376 390–5 415 856.

The AMR determinants harboured by the MDR isolates belonging to Cluster 2 were also co-located and conferred resistance to aminoglycosides, ampicillin, trimethoprim, sulphonamides and chloramphenicol. However, the specific AMR genes were different *bla*_{TEM-1}/*strA*, *strB/aadA-1b* *aac(3)-Iva/aph(4)-Ia*, *aadD/dfrA-1/sul-1/sul-2/tetA/floR*, and, in contrast to MDR isolates belonging to 2.8.351%, they were located on a IncHI2 plasmid.

Long-read sequencing data from isolates belonging to the clade designated Cluster 3 identified two groups of AMR determinants, both located on the same IncHI2 plasmid. Group 1 comprised *bla*_{TEM-1}/*strA*, *strB/aadA-1/dfrA-1/sul-1/sul-2/tetA*, and Group 2 had *strA*, *strB/sul-2/floR* and *bla*_{CTX-M-15}/*qnrS-1*.

Most MDR isolates in the clade designated Cluster 4 had the AMR profile *bla*_{TEM-1}/*strA*, *strB/dfrA-8/sul-2/tetA*. All AMR genes were co-located on a 26 887 bp fragment of DNA integrated into the Locus of Enterocyte Effacement on the chromosome. A small subset of isolates also had an IncHI2 plasmid encoding *ermB/mph-B/strA*, *strB/aadA-1/aadA-2/aadA-8b/aac(3)-Iva/aph(4)-Ia*, *aadD/dfrA-1/dfrA-12/sul1/sul2/tetA*.

Travel history and association with sub-lineages

As shown in previous studies^{17,19}, the analysis of the isolates in this study indicated that the UK sub-lineages Ic and IIb were mostly associated with domestically acquired infection with only 15/342 (4.4%) and 9/215 (4.2%) of cases, respectively, reporting travel outside the UK before onset of symptoms (Table 1). By contrast, a higher proportion of cases infected with isolates each belonging to sub-lineages Ia (11/38, 28.9%), Ib (18/64, 28.1%), IIa (86/323, 26.6%) and IIc (93/440, 21.1%) reported travelling outside the UK within 7 days of onset of symptoms, and were designated travel associated cases (Table 1, Supplementary Table S1).

Discussion

Since the 1990s, surveillance of STEC O157:H7 in England has included monitoring AMR profiles to assess the risk of transmission of MDR pathogens from animals to humans via the food chain.^{7,38,39} In this study, we reviewed the AMR gene profiles of STEC O157:H7 by sub-lineage and found that AMR genes were unevenly distributed across the phylogeny. AMR determinants were identified in less frequently in sub-lineages Ia, Ic, IIa and IIb, whereas sub-lineages Ib, IIa and IIc, and lineage I/II comprised a higher proportion of isolates that were predicted to be resistant to at least one class of antimicrobial based on the AMR gene content.

Over the last two decades, the proportion of isolates of STEC O157:H7 harbouring resistance to at least one class of antibiotic has decreased from 20% to 14.7%.^{7,38} Our previous analysis of the evolutionary history of STEC O157:H7 in the UK showed that during the 1980s and 1990s lineage I/II was the dominant UK lineage.^{17,18} Lineage I/II comprises a higher proportion of isolates harbouring AMR determinants than sub-lineages Ic or IIb. The decrease in the number of cases infected with STEC O157:H7 belonging to I/II, and the emergence of sub-lineage Ic, and more recently IIb, may explain the decreasing proportion of strains exhibiting AMR. Historically, and in this study, most lineage I/II isolates that had AMR determinants were predicted to be resistant to streptomycin, tetracycline and the sulphonamides (STR/TET/SUL).^{38,39} STEC O157:H7 is zoonotic and endemic in the UK cattle and sheep population, and the prevalence of genes encoding resistance to STR/TET/SUL in lineage I/II is consistent with these antimicrobials being used as therapeutic options in veterinary practice.⁷ Improvements in the animal husbandry practices in the UK and more regulated use of these antimicrobials since the 1990s resulting in the reduction of selective pressure, may

explain why strains belonging to the more recently emerged sub-lineages Ic and IIb have remained, for the most part, fully susceptible to the classes of antibiotics included in this study.

In contrast to sub-lineages Ic and IIb, there is little or no evidence that sub-lineages Ia and Ib are endemic in the UK ruminant population, and infection is most likely to occur either following travel outside the UK or via the consumption of imported contaminated food. Sub-lineage Ib has a higher proportion of isolates encoding AMR genes than sub-lineage Ia, even though both sub-lineages appear to be widely distributed across all regions of the globe, and the reasons for this difference remain unclear.

We have previously presented evidence that, within sub-lineages IIa and IIc isolated from human cases in the UK, there are clades that are endemic in the UK cattle population and those that are imported into the UK, either by travellers returning to the UK from abroad or via non-domestically produced food items.^{19,40} Sub-lineage IIc has a higher proportion of isolates encoding AMR genes than sub-lineage IIa, even though both sub-lineages comprise a similar proportion of travel related cases. As before, the reasons for the different proportions of AMR isolates between the two sub-lineages are unclear. It is possible that strains from certain sub-lineages originate from regions where unregulated antibiotic use in animals and/or treatment of human infection is contributing to the persistence and transmission of AMR in the local STEC population. Alternatively, certain sub-lineages may be better adapted to acquiring and maintaining MGE encoding AMR.

As well as being unevenly distributed across the different sub-lineages of STEC O157:H7, AMR genes are unevenly distributed between different clades within the same sub-lineage.⁴⁰ Most AMR determinants were co-located either on the chromosome and/or on large plasmids. Dallman *et al.*⁴⁰ concluded that domestic clades of sub-lineage IIc were largely populated with susceptible isolates, whereas non-domestic clades contained a higher proportion of MDR isolates. Once stabilized in the population, these clades may persist in certain environments where antibiotic use is high. AMR determinants located on MGE may transfer from pathogen to commensal bacteria in the gut, increasing the pool of MDR bacteria that may cause life-threatening extraintestinal infections. Monitoring AMR in gastrointestinal pathogens may provide an early warning of emerging risks to public health regarding the clinical management and empirical treatment of infectious diseases.

Resistance mechanisms not detected in our previous study in 2016⁷ were identified, although the numbers of isolates that had these resistance mechanisms were low. We identified mutations in *gyrA* and *parC* predicted to exhibit an increase in MIC to fluoroquinolones, *bla*_{CTX-M-15} predicted to confer resistance to the third-generation cephalosporins, and *mphA* and *ermB* predicted to confer resistance to azithromycin. The genes conferring resistance to the third-generation cephalosporins and azithromycin were plasmid-encoded, and therefore had the potential to be transferred to other bacteria in the gut.^{9,41,42}

It has been suggested that antimicrobial use in food-producing animals may be a risk factor in the transmission of MDR strains of *E. coli* from animals to humans, via the food chain. Analysis of data from outbreak investigations has provided evidence of this transmission route,⁴³ however, it has been

challenging to identify associations between consumption of contaminated food and infection with MDR STEC O157:H7 in sporadic cases. Moving forward, real-time WGS analysis of genome derived AMR determinants, including characterization of drug determinant regions, will enable us to monitor the persistence of MDR strains in the animal reservoir and assess the risk factors associated with transmission to humans. Wider application of long-read sequencing for public health surveillance will enable us to observe the loss and acquisition of MGE encoding AMR determinants across the STEC O157:H7 phylogeny and, potentially, transmission to other bacteria in the gut of both humans and animals, and in the environment.

In summary, the majority of STEC O157:H7 causing domestically acquired infection in the UK are susceptible to most classes of antimicrobials, whereas strains associated with travellers' diarrhoea are more likely to be multidrug resistant and may exhibit resistance to ampicillin, trimethoprim and chloramphenicol. Historically, in the UK the unregulated use of antibiotics in veterinary practice may have been a selective pressure for the acquisition of resistance genes, however, in this dataset most MDR isolates causing human infection on the UK were acquired abroad. The implementation of long-read sequencing for routine surveillance will enable us to monitor the emergence and spread of both MDR enteric pathogens and the AMR genes, thus enabling us to track the transmission of AMR at the pathogen and mobile genetic element level.

Funding

David Greig and Claire Jenkins are affiliated to the National Institute for Health and Care Research Health Protection Research Unit (NIHR HPRU) in Gastrointestinal Infections at University of Liverpool in partnership with the United Kingdom Health Security Agency (UKHSA), and the University of Warwick. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR, the Department of Health and Social Care, or UKHSA.

Transparency declarations

Nothing to declare.

Supplementary data

Figure S1 and Table S1 are available as [Supplementary data](#) at JAC Online.

References

- 1 Baker KS, Dallman TJ, Ashton PM *et al.* Intercontinental dissemination of azithromycin-resistant shigellosis through sexual transmission: a cross-sectional study. *Lancet Infect Dis* 2015; **15**: 913–21. [https://doi.org/10.1016/S1473-3099\(15\)00002-X](https://doi.org/10.1016/S1473-3099(15)00002-X)
- 2 Ingle DJ, Nair S, Hartman H *et al.* Informal genomic surveillance of regional distribution of *Salmonella* Typhi genotypes and antimicrobial resistance via returning travellers. *PLoS Negl Trop Dis* 2019; **13**: e0007620. <https://doi.org/10.1371/journal.pntd.0007620>
- 3 Bardsley M, Jenkins C, Mitchell HD *et al.* Persistent transmission of shigellosis in England is associated with a recently emerged multidrug-resistant strain of *Shigella sonnei*. *J Clin Microbiol* 2020; **58**: e01692-19. <https://doi.org/10.1128/JCM.01692-19>

- 37** Gilchrist CLM, Chooi YH. Clinker & clustermap.js: automatic generation of gene cluster comparison figures. *Bioinformatics* 2021; **37**: 2473–5. <https://doi.org/10.1093/bioinformatics/btab007>
- 38** Willshaw GA, Cheasty T, Smith HR et al. Verocytotoxin-producing *Escherichia coli* (VTEC) O157 and other VTEC from human infections in England and Wales: 1995–1998. *J Med Microbiol* 2001; **50**: 135–42. <https://doi.org/10.1099/0022-1317-50-2-135>
- 39** Threlfall EJ, Ward LR, Frost JA et al. The emergence and spread of antibiotic resistance in food-borne bacteria. *Int J Food Microbiol* 2000; **62**: 1–5. [https://doi.org/10.1016/S0168-1605\(00\)00351-2](https://doi.org/10.1016/S0168-1605(00)00351-2)
- 40** Dallman TJ, Jalava K, Verlander NQ et al. Identification of domestic reservoirs and common exposures in an emerging lineage of Shiga toxin-producing *Escherichia coli* O157:H7 in England: a genomic epidemiological analysis. *Lancet Microbe* 2022; **3**: e606–15. [https://doi.org/10.1016/S2666-5247\(22\)00089-1](https://doi.org/10.1016/S2666-5247(22)00089-1)
- 41** Mitchell HD, Whitlock G, Zdravkov J et al. Prevalence and risk factors of bacterial enteric pathogens in men who have sex with men: a cross-sectional study at the UK's largest sexual health service. *J Infect* 2023; **86**: 33–40. <https://doi.org/10.1016/j.jinf.2022.10.033>
- 42** Thorley K, Charles H, Greig DR et al. Emergence of extensively drug-resistant and multidrug-resistant *Shigella flexneri* serotype 2a associated with sexual transmission among gay, bisexual, and other men who have sex with men, in England: a descriptive epidemiological study. *Lancet Infect Dis* 2023; **23**: 732–9. [https://doi.org/10.1016/S1473-3099\(22\)00807-6](https://doi.org/10.1016/S1473-3099(22)00807-6)
- 43** Kaindama L, Jenkins C, Aird H et al. A cluster of Shiga Toxin-producing *Escherichia coli* O157:H7 highlights raw pet food as an emerging potential source of infection in humans. *Epidemiol Infect* 2021; **149**: e124. <https://doi.org/10.1017/S0950268821001072>