MICROBIAL GENOMICS

OUTBREAK REPORT

Greig *et al.*, *Microbial Genomics* 2021;7:000545

DOI 10.1099/mgen.0.000545

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. Gharbia[1] 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: [1]National Infection Service, Public Health England, London, NW9 5EQ, UK; [2]Division 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 bacteriophag 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.

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 Ic [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.

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.

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*) 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 Revolugen Fire Monkey DNA extraction kit (Revolugen, 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

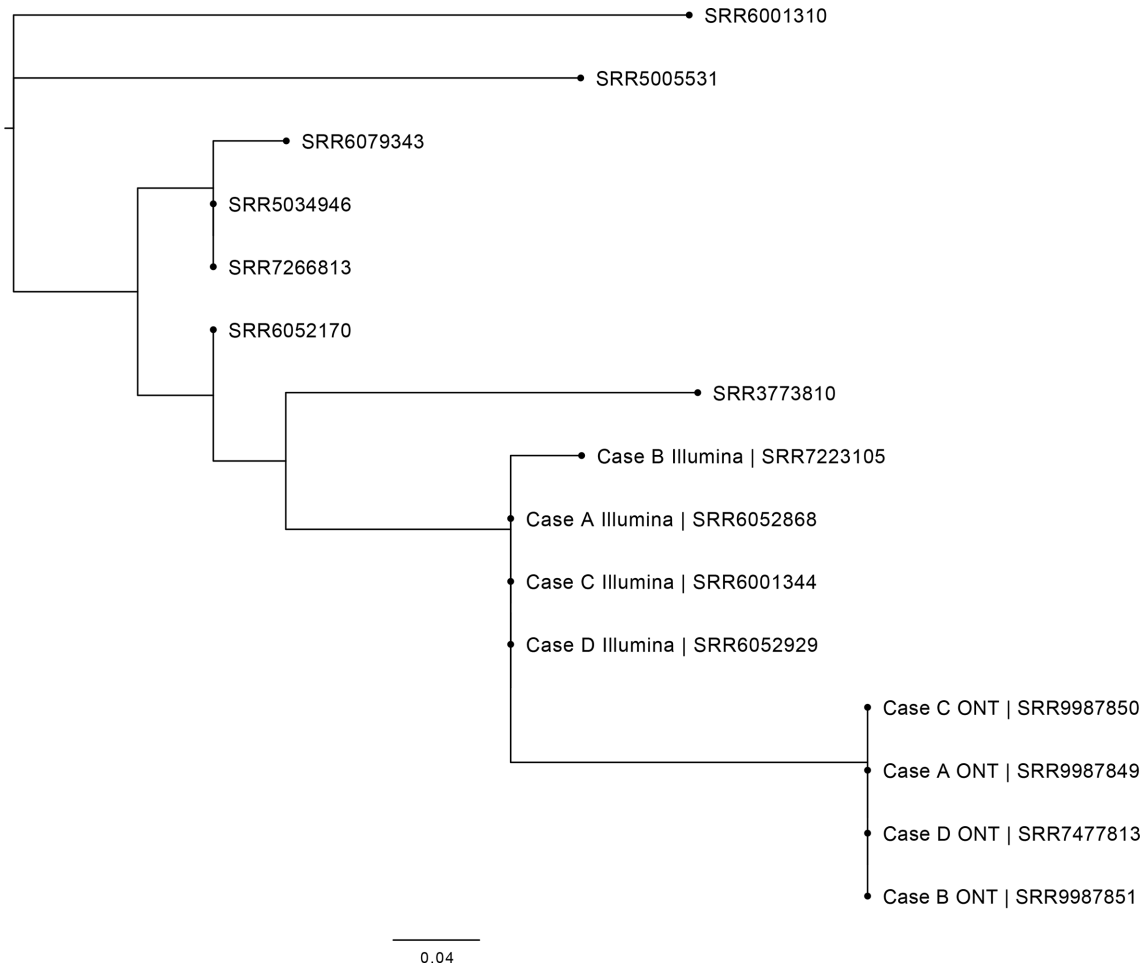

Fig. 1. Maximum-likelihood phylogenetic tree of the four outbreak samples within this outbreak and the seven most closely related isolates in the PHE archive to show context. The four outbreak cases (A–D) sequenced by both Illumina and Nanopore technologies are shown.

Technologies) into FASTQ format and grouped in each samples' respective barcode. Samples were re-demultiplexed using Deepbinner 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=dcm,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 technolgoies 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 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 V0.11.1 [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, *sbcB*, 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 Ic and Ia, 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	*lexA*	*aphA*	22035	22035	22034	22036	/	403708–425743	403729–425764	403712–425746	403708–425744
2*	*tRNA-Thr* (cgt)†	*prgR*	26826	26823	26823	26821	/	1101413–1128239	1101440–1128263	1101421–1128244	1101414–1128235
3	*ybhC*†	*ybhB*	38095	38095	38096	38095	Sp3 (99%/100%)	1697277–1735372	1697420–1735515	1697397–1735493	1697388–1735483
4	*yccA*†	*tRNA-Ser* (tga)	48257	48277	48265	48263	Sp4 (99%/100%)	1967370–2015627	1967513–2015790	1967491–2015756	1967481–2015744
5	*potC*†	*potB*	47889	47679	47879	47896	Sp6 (99%/100%)	2283891–2331780	2284054–2331733	2284019–2331898	2284012–2331908
6	*roxA*†	*phoQ*	10436	10436	10436	10436	/	2337022–2347458	2336975–2347411	2337140–2347576	2337150–2347586
7	*icd*†	*caeB*	40992	40991	40992	40991	Sp8 (99%/100%)	2356230–2397222	2356183–2397174	2356,348–2397340	2356358–2397349
8*	*ompW*†	*rspR*	145151	144981	145040	53636	/	2495566–2640717	2495518–2640499	2495685–2640725	2495693–2549329
A	*hipA*	*ydeP*	/	/	/	28513	/	/	/	/	2589222–2617735
9	*trpA*	*rspA*	59339	59340	59322	59863	/	2918047–2977386	2917834–2977174	2918061–2977383	2855169–2915032
10	*yodB*	*tRNA-Ser* (cga)†	43681	43694	43684	43690	Sp14 (99%/98%)	3370042–3413723	3369847–3413541	3370056–3413740	3307170–3350860
11 (*stx2c*)	*yeeA*	*tnpA*†	55650	56910	55652	55655	/	3457661–3513311	3456222–3513132	3457677–3513329	3394797–3450452
12	*yehV*	*yehV*†	45720	45725	45732	45719	/	3654766–3700486	3654584–3700309	3654782–3700514	3591904–3637623
13 (*stx2a*)	*yfdC*	*argW*†	60239	60237	60239	60242	/	3948563–4008802	3948386–4008623	3948591–4008830	3885700–3945942
14	*argW*†	*lacY*	8233	8233	8233	8233	Sp16 (100%/100%)	4009234–4017467	4009055–4017288	4009262–4017495	3946374–3954607
15	*ssrA*	*alpA*†	22107	22103	22098	22106	Sp17 (99%/99%)	4294,770–4316877	4,294,584–4316687	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*	*yeeB*	15044	15045	15044	15046	SpLE2 (99%/100%)	3439339–3454383	3439883–3454928	3439719–3454763	3377102–3392148
PLE5	tRNA-*Phe*	*pitB*	23334	23333	23333	23332	SpLE3 (99%/100%)	4664210–4687544	4664345–4687678	4664588–4687921	4601991–4625323
PLE6	*selC*	*yicL*	50390	50389	50390	50390	SpLE4 (99%/79%)	5392598–5442988	5392972–5443361	5393215–5443605	5330617–5381007

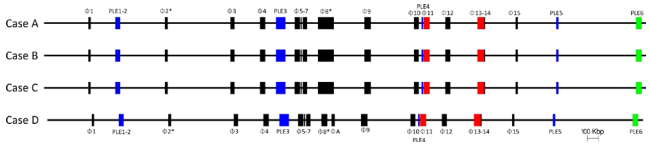

Fig. 2. The relative positions of the 15 prophages and six prophage-like elements within the chromosomes of samples A–C, 16 prophages within sample D (order descending). Red prophages indicate *stx* gene encoding prophages. Blue indicates prophage like elements. Green indicates the locus of enterocyte 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).

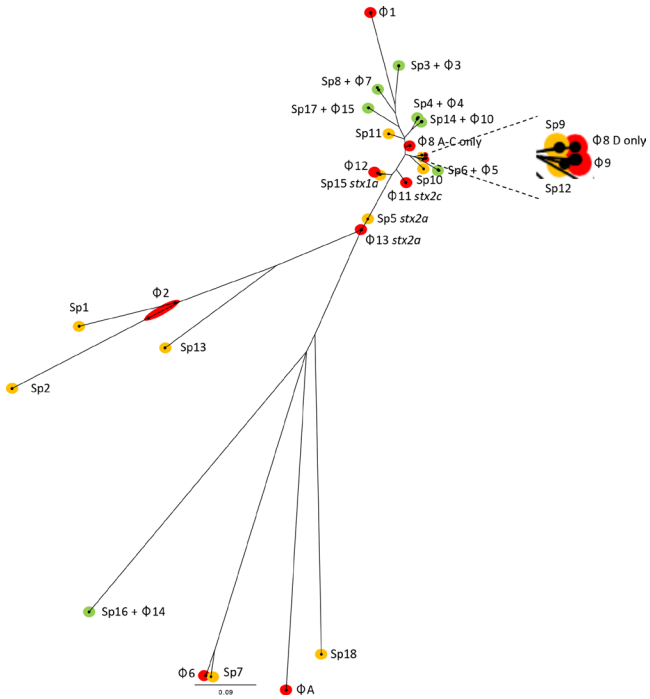

Fig. 3. Mid-rooted neighbour-joining tree of Jaccard distances showing prophages from cases A–D with prophages from BA000007 (Sakai). Grouped by green; prophages shared in cases A–D and Sakai, yellow; Sakai only and red; prophages unique to cases A–D.

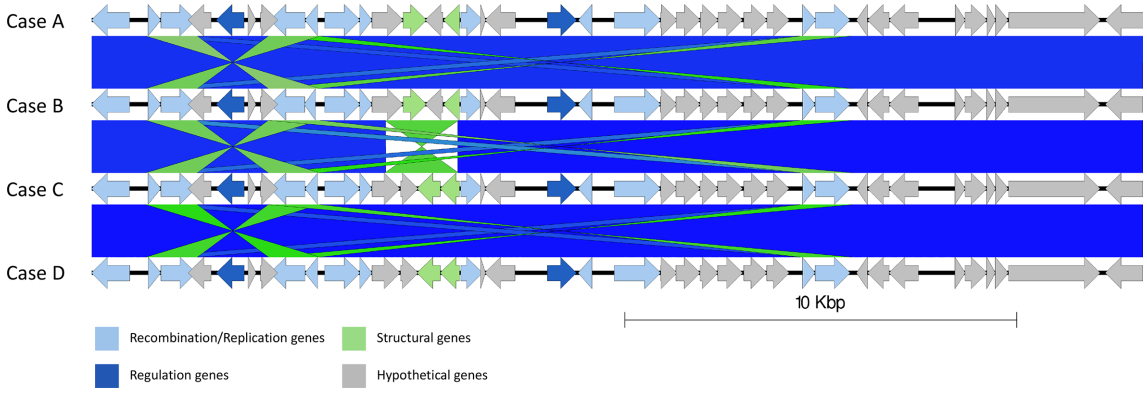

Fig. 4. Easyfig alignment of prophage 2 in all cases order descending A–D, detailing a small inversion from samples A and B relative to samples C and D (light blue). Arrow indicates gene direction. Recombination/replication genes shown in light blue, regulation associated genes in dark blue. Structure 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 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.

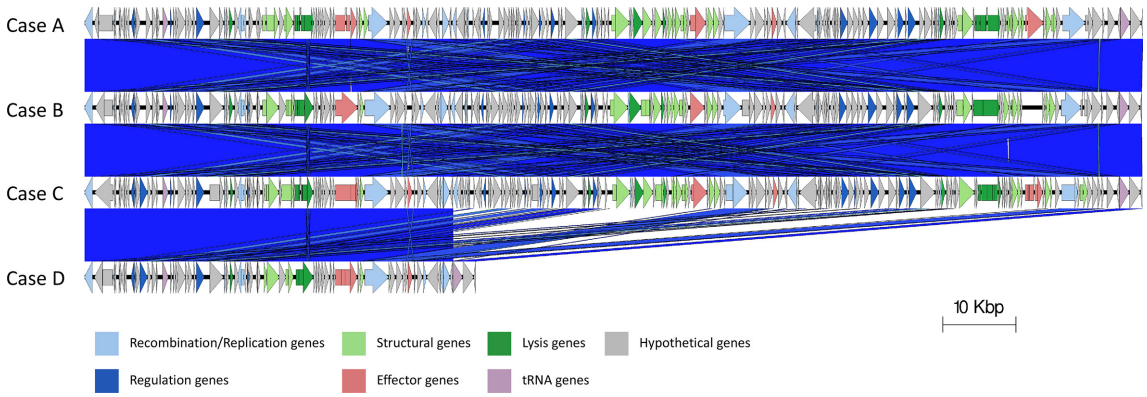

Fig. 5. Easyfig alignment of prophage 8 in all cases order descending A–D. Showing a substitution in sample D. Arrow indicates gene direction. Recombination/replication genes shown in light blue, regulation associated genes in dark blue. Effector genes are shown in pink, structure and lysis associated 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 O57: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

1. 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.
2. 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.
3. 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.
4. 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.
5. 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.
6. 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.
7. 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.
8. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics* 2014;30:2114–2120.
9. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* 2010;26:589–595.
10. 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.
11. 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.
12. 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.
13. Wick RR, Judd LM, Holt KE. Deepbinner: demultiplexing barcoded Oxford nanopore reads with deep convolutional neural networks. *PLoS Comput Biol* 2018;14:e1006583.
14. 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.
15. Wick RR. Porechop. https://github.com/rrwick/Porechop
16. Wick RR. Filtlong. https://github.com/rrwick/Filtlong
17. Kolmogorov M, Yuan J, Lin Y, Pevzner PA. Assembly of long, error-prone reads using repeat graphs. *Nat Biotechnol* 2019;37:540–546.
18. Loman NJ, Quick J, Simpson JT. A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nat Methods* 2015;12:733–735.
19. 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.
20. 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.
21. 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.
22. 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.
23. 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.