MICROBIAL GENOMICS

RESEARCH ARTICLE

Yara *et al.*, *Microbial Genomics* 2020;6

DOI 10.1099/mgen.0.000334

MICROBIOLOGY SOCIETY

OPEN DATA OPEN MICROBIOLOGY

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. Gally[3], 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: [1]Norwich Medical School, University of East Anglia, Norwich, UK; [2]National Infection Service, Public Health England, London NW9 5EQ, UK; [3]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: *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.

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.

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.

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 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 QIAsymphony 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	*stx1a* prophage SBI and size (bp)	*stx2a* prophage SBI and size (bp)	*stx2c* prophage SBI and size (bp)	Reference	BioProject accession no.	Assembly accession no.
PHE archive											
E30228	Ia	PT4	Stx1a/2a	5 416 109	15	*yehV* (47 594)	*wrbA* (62 890)	–	[41]	PRJNA315192	VXJO00000000
E34500	I/IIa	PT2	Stx2a/2c	5 359 964	14	–	*argW* (62 149)	*sbcB* (57 463)	[2]	PRJNA315192	VXJN00000000
E45000	I/IIb	PT49	Stx2a	5 386 698	17	–	*sbcB* (44 014)	–	This study	PRJNA315192	VXJM00000000
E116508	Ic	PT21/28	Stx2a/2c	5 571 891	17	–	*argW* (71 870)	*sbcB* (59 105)	This study	PRJNA315192	VXJP00000000
315176	IIb	PT8	Stx2a	5 579 120	16	–	*sbcB* (61 851)	–	[10]	PRJNA315192	VXJQ00000000
267849	IIa	PT34	Stx2a/2c	5 510 912	16	–	*yecE* (47 242)	*sbcB* (61 840)	[43]	PRJNA315192	VXJR00000000
Publicly available											
9000	Ic	PT21/28	Stx2a/2c	5 516 497	17	–	*argW* (65 158)	*sbcB* (57 408)	[33]	PRJNA336330	CP018252
397404	Ic	PT21/28	Stx2a/2c	5 618 435	13	–	*argW* (70 472)	*sbcB* (59 098)		PRJNA315192	CP043019
155	Ic	PT32	Stx2a	5 513 008	18	–	*yecE* (50 015)	–	[33]	PRJNA336330	CP018237
350	IIc	PT8	Stx1a/2c	5 411 823	16	*yehV* (49 867)	–	*sbcB* (57 747)	[11]	PRJNA336330	CP018243
272	I/IIa	PT2	Stx2a	5 474 193	16	–	*argW* (65 675)	–	[13]	PRJNA336330	CP018239
644	IIc	PT8	Stx1a/1a/2c	5 831 209	18	*yehV* (49 544) *argW* (64 569)	–	*sbcB* (58 210)	[12]	PRJNA321984	CP015831
180	IIc	PT54	Stx1a/1a/2c	5 509 528	15	*yehV* (49 544) *argW* (64 569)	–	*sbcB* (61 558)	[12]	PRJNA321984	CP015832
Sakai	Ia	–	Stx1a/2a	5 498 450	18	*yehV* (47 650)	*wrbA* (62 142)	–	[16]	PRJNA57781	NC_002695
EDL933	Ia	–	Stx1a/2a	5 547 323	14	*yehV* (47 596)	*wrbA* (61 066)	–	[14]	PRJNA253471	CP008957
EC4115	I/IIa	–	Stx2a/2c	5 572 075	17	–	*argW* (71 540)	*sbcB* (60 476)	[15]	PRJNA224116	NC_011353
TW14359	I/IIa	–	Stx2a/2c	5 528 136	17	–	*argW* (71 540)	*sbcB* (60 476)	[15]	PRJNA224116	NC_013008

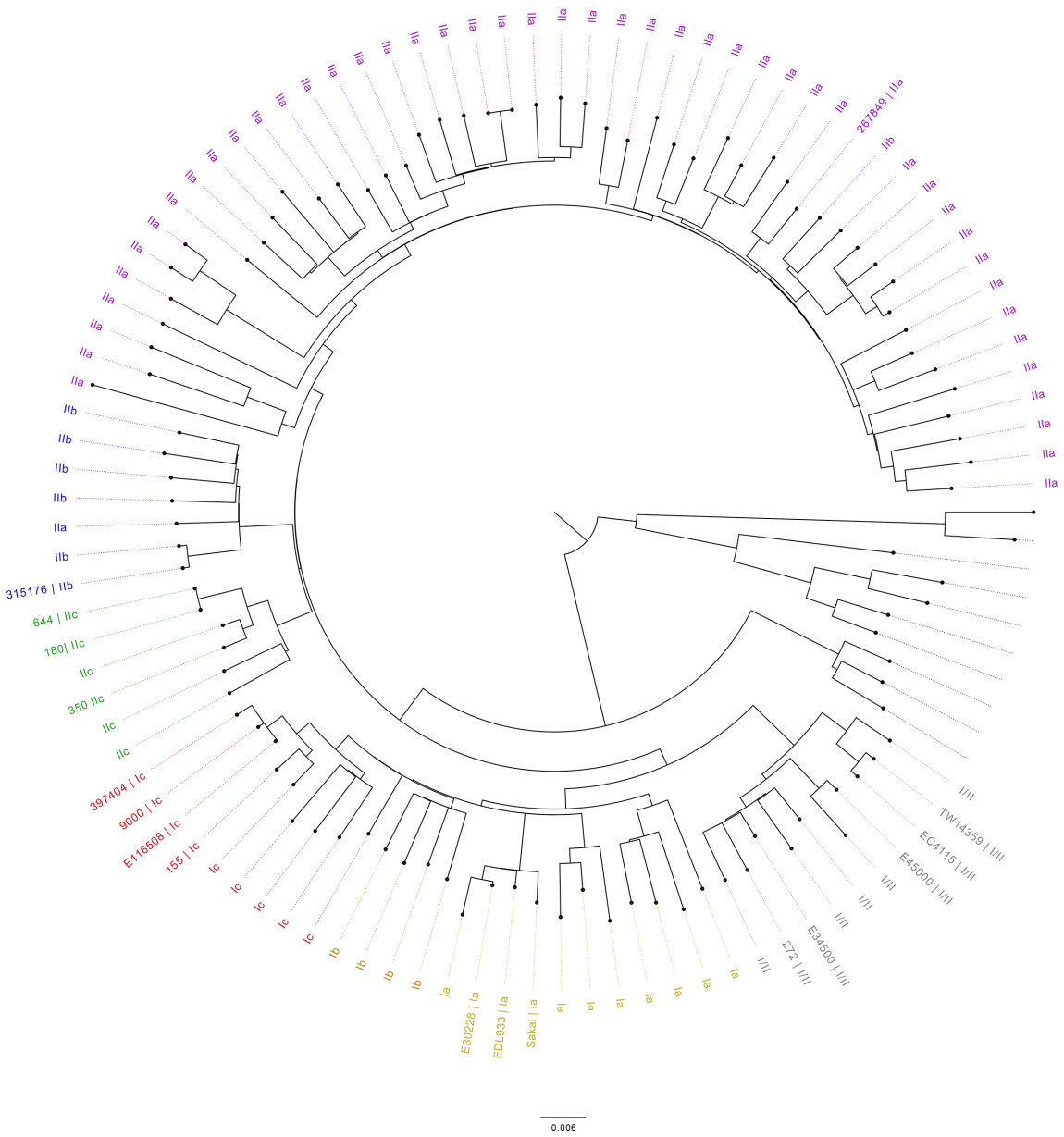

Fig. 1. Maximum-likelihood phylogenetic tree of 105 genomes including the 17 (labelled) publicly available genomes and nanopore sequenced genomes produced during this study. Sub-lineages are coloured as follows: I/II, grey; Ia, yellow; Ib, orange; Ic, red; IIa, purple; IIb, blue; IIc, green. Scale bar indicates kbp.

Technologies) into FASTQ format and grouped in each samples' respective barcode. De-multiplexing was performed using Deepbinner v0.2.0 [18]. Run metrics were generated using Nanoplot v1.8.1 [19]. The barcode and y-adapter 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

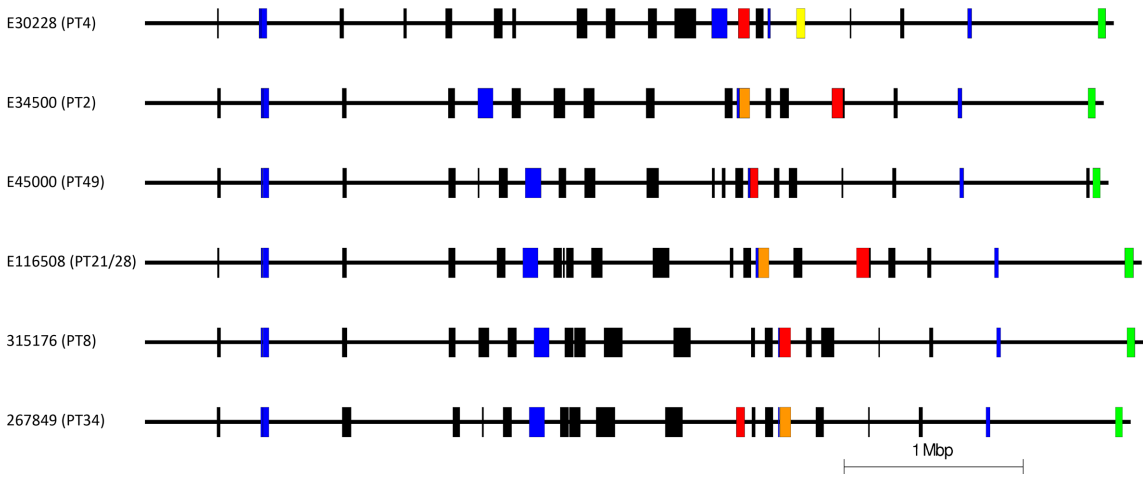

Fig. 2. Easyfig diagram representing the chromosome and prophage content within the samples sequenced in this study (in descending order PT4, PT2, PT49, PT21/28, PT8 and PT34). *stx2a*-encoding, *stx2c*-encoding and *stx1*-encoding prophages are highlighted in red, orange and yellow, respectively. Non-*stx*-encoding 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=dcm 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 v 1.13 [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 though 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, VXJO00000000; E34500, VXJN00000000; E45000, VXJM00000000; E116508, VXJP00000000; 315176, VXJQ00000000; and 267849, VXJR00000000.

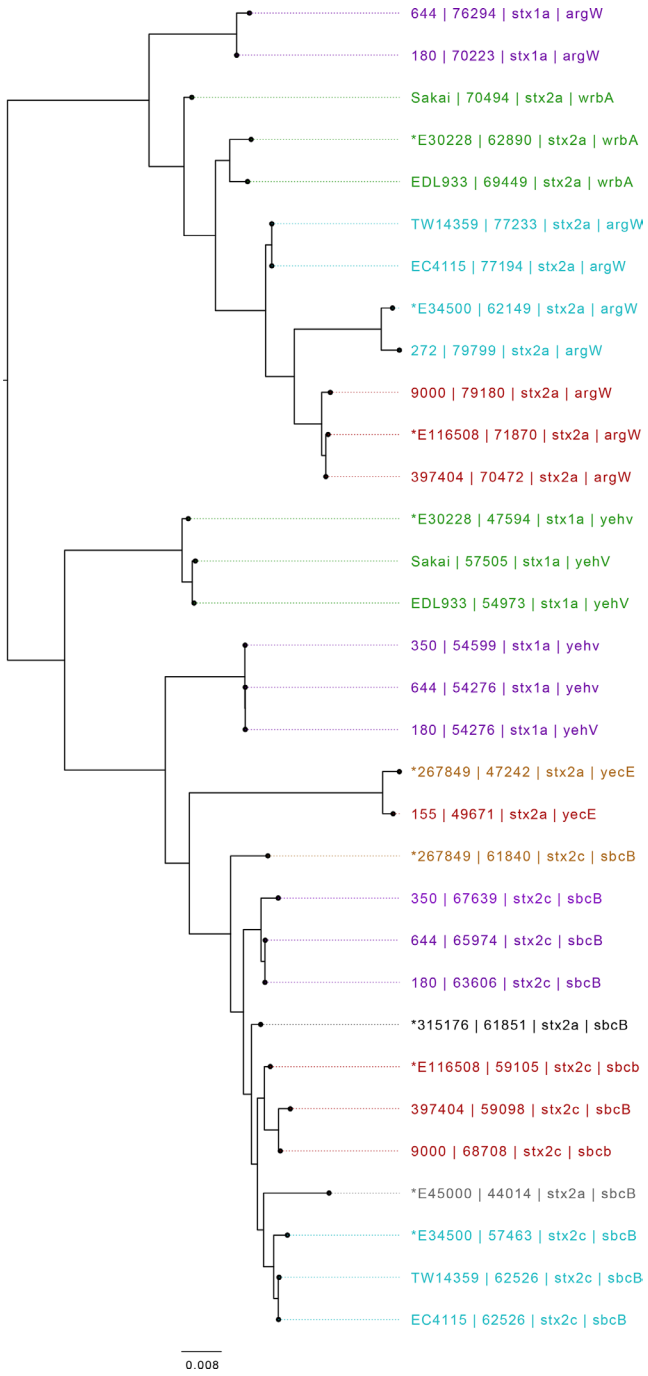

Fig. 3. Mid-rooted tree of *stx*-encoding prophages based on Jaccard distance produced from Mash. Strains are annotated with strain ID, length (bp), *stx* profile and SBI. Strains sequenced during this study have prophages that are hown preceded by an *. Strains are coloured by sub-lineage: green, Ia; 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 5 579 120 bp, 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

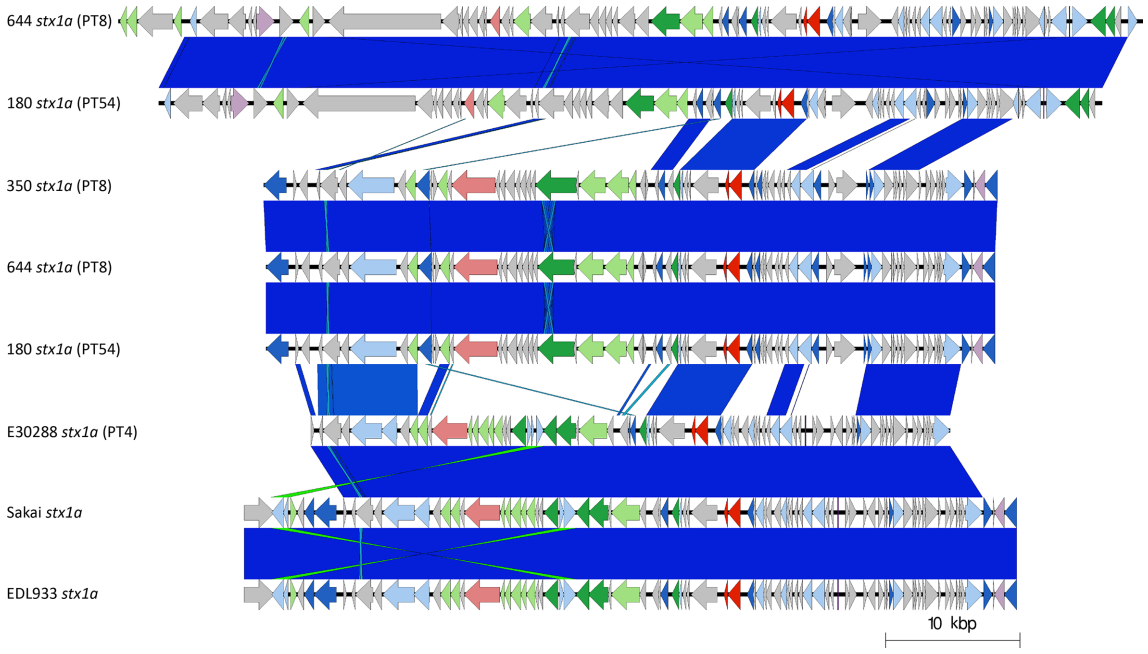

Fig. 4. Easyfig plot comparing the *stx1a*-encoding prophages from 644 (×2), 180 (×2), 350, E30288, EDL933 and Sakai. Arrows indicate gene directions. *stx* genes are shown in red; recombination/replication genes are shown in light blue; regulation-associated genes are shown in dark blue; effector genes 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).

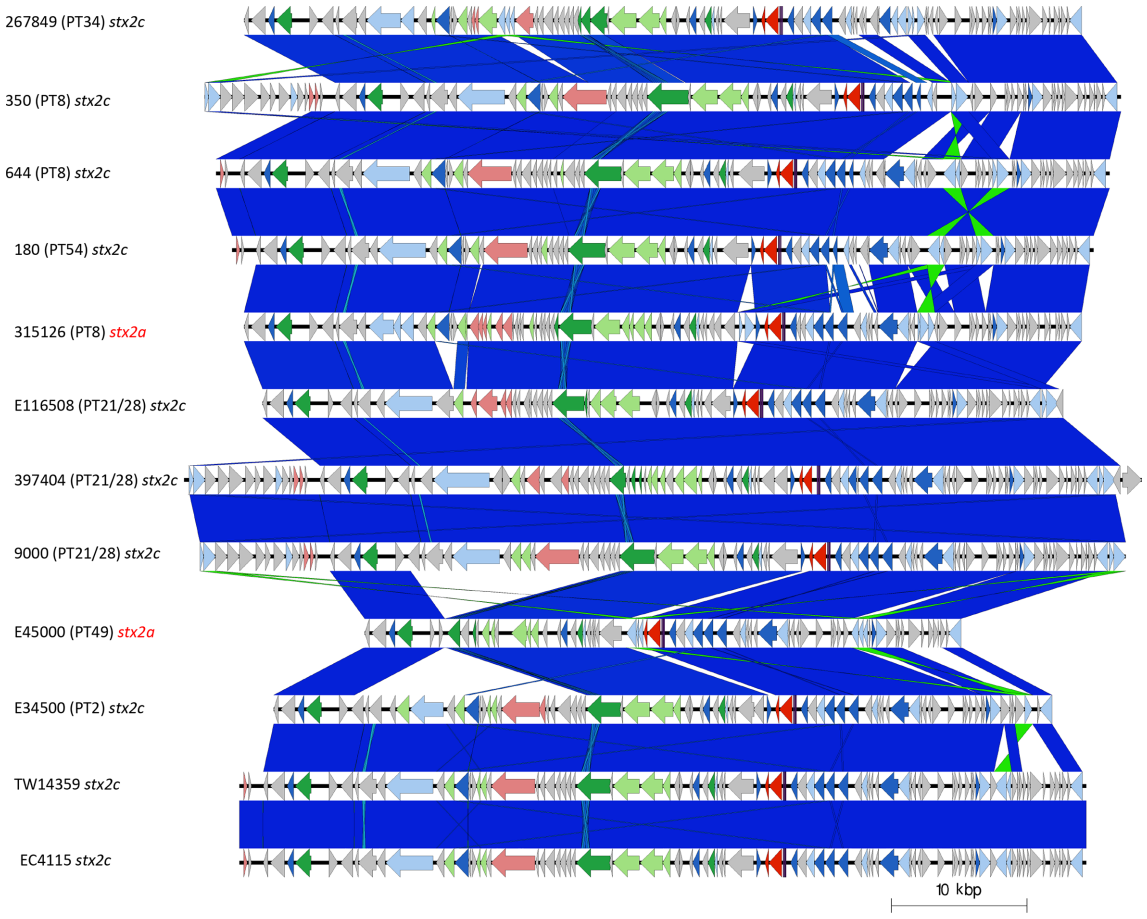

Fig. 5. Easyfig plot comparing the *stx2c*-encoding prophages from all samples in the study, including two *stx2a* prophages that are in a *stx2c*-associated prophage structure (315126 and E45000). Arrows indicate gene directions. *stx* genes are shown in red; recombination/replication genes are shown 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.

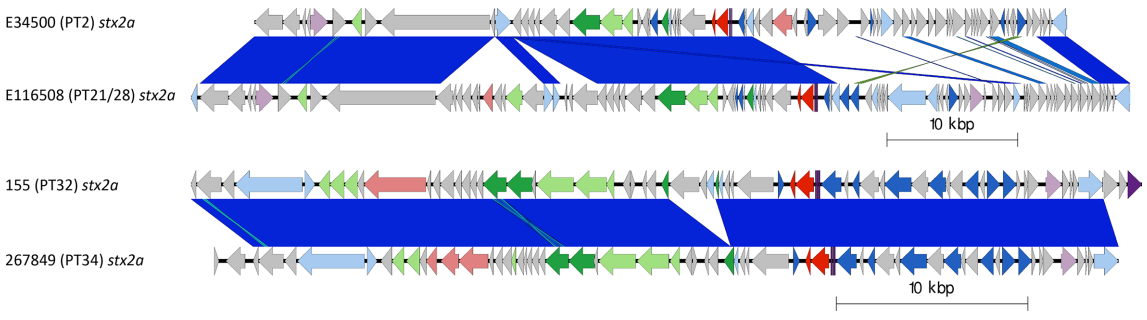

Fig. 6. Two Easyfig plots comparing the *stx2a*-encoding prophages from E45000 with E116508 (above) and 155 and 267849 (below), in descending order. Arrows indicate gene directions. *stx* genes are shown in red; recombination/replication genes shown in light blue; regulation-associated genes are shown 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 *sbcB* 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, VXJO00000000–VXJR00000000 (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, Petrovska 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* 0157: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 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.
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 VTl 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.