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

David R. Greig1,2, Amy F. W. Mikhail1, Timothy J. Dallman1,2 and Claire Jenkins1*

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

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/IA, and I/IB) (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 I/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) (Scheutz 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 Timmomatic 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 v1.0 (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>
<thead>
<tr>
<th>Sub-lineage</th>
<th>Total number of isolates</th>
<th>Number of sporadic isolates</th>
<th>Number of clusters detected at the 250 SNP level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ia</td>
<td>31</td>
<td>29</td>
<td>9</td>
</tr>
<tr>
<td>Ib</td>
<td>25</td>
<td>25</td>
<td>4</td>
</tr>
<tr>
<td>Ic</td>
<td>234</td>
<td>183</td>
<td>4</td>
</tr>
<tr>
<td>IIA</td>
<td>325</td>
<td>98</td>
<td>28</td>
</tr>
<tr>
<td>IIB</td>
<td>139</td>
<td>44</td>
<td>4</td>
</tr>
<tr>
<td>IIC</td>
<td>285</td>
<td>246</td>
<td>1</td>
</tr>
<tr>
<td>I/II</td>
<td>37</td>
<td>24</td>
<td>3</td>
</tr>
<tr>
<td>NSF</td>
<td>6</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>SF</td>
<td>21</td>
<td>18</td>
<td>5</td>
</tr>
<tr>
<td>Total</td>
<td>1103</td>
<td>671</td>
<td></td>
</tr>
</tbody>
</table>
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 2** | Age and sex ratio and characteristics of isolates from UK travelers returning from the ISC and Uganda 2006–2016 (number in parentheses).

<table>
<thead>
<tr>
<th>Country</th>
<th>Total</th>
<th>Male</th>
<th>Female</th>
<th>Adult</th>
<th>Child</th>
<th>Lineage (n)</th>
<th>Stx profile (n)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bangladesh</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>IIa (4)</td>
<td>Stx2c (4)</td>
</tr>
<tr>
<td>Bhutan</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>Ia (1)</td>
<td>Stx2c (1)</td>
</tr>
<tr>
<td>India</td>
<td>32</td>
<td>19</td>
<td>16</td>
<td>24</td>
<td>11</td>
<td>IIa (32)</td>
<td>Stx2c (32)</td>
</tr>
<tr>
<td>Maldives</td>
<td>0</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Nepal</td>
<td>0</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Pakistan</td>
<td>16</td>
<td>6</td>
<td>9</td>
<td>2</td>
<td>14</td>
<td>la (1)</td>
<td>Stx1a/Stx2a</td>
</tr>
<tr>
<td>Sri Lanka</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>IIa (3)</td>
<td>Stx2c (3)</td>
</tr>
<tr>
<td>Total ISC</td>
<td>56</td>
<td>29</td>
<td>29</td>
<td>31</td>
<td>28</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Uganda</td>
<td>8</td>
<td>2</td>
<td>6</td>
<td>6</td>
<td>2</td>
<td>IIa (7)</td>
<td>Stx2a/stx2c (4)</td>
</tr>
</tbody>
</table>

**TABLE 3** | Size, position, and integration sites of all 19 prophages (P) and six prophage-like elements (PLE) within sample 194195.

<table>
<thead>
<tr>
<th>Prophage detected by PHASTER</th>
<th>Gene 5’ to prophage</th>
<th>Gene 3’ to prophage</th>
<th>Size (bp)</th>
<th>Position</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>cpxP</td>
<td>fieF</td>
<td>31,356</td>
<td>237,188–268,544</td>
</tr>
<tr>
<td>P2a</td>
<td>zur</td>
<td>aphA</td>
<td>19,508</td>
<td>433,909–453,417</td>
</tr>
<tr>
<td>PLE1</td>
<td>leuX</td>
<td>HypotheticalP</td>
<td>9,862</td>
<td>678,235–688,097</td>
</tr>
<tr>
<td>PLE2</td>
<td>intA</td>
<td>nanM</td>
<td>33,695</td>
<td>688,557–722,252</td>
</tr>
<tr>
<td>P3a</td>
<td>tRNA-Thr (cgt)</td>
<td>prgF</td>
<td>54,474</td>
<td>1,128,970–1,183,444</td>
</tr>
<tr>
<td>P4</td>
<td>ybhC</td>
<td>ybhB</td>
<td>38,693</td>
<td>1,751,265–1,789,958</td>
</tr>
<tr>
<td>P5</td>
<td>yccaC</td>
<td>tRNA-Ser (tga)</td>
<td>49,561</td>
<td>2,021,966–2,071,527</td>
</tr>
<tr>
<td>PLE3</td>
<td>ycdU</td>
<td>tRNA-Ser (gga)</td>
<td>85,778</td>
<td>2,170,090–2,255,868</td>
</tr>
<tr>
<td>P8a</td>
<td>potC</td>
<td>potB</td>
<td>45,568</td>
<td>3,160,400–3,205,968</td>
</tr>
<tr>
<td>P10a</td>
<td>ttcA</td>
<td>rpoB</td>
<td>101,788</td>
<td>3,617,610–3,661,297</td>
</tr>
<tr>
<td>P11</td>
<td>yehV</td>
<td>mlrA</td>
<td>52,982</td>
<td>3,900,672–3,953,654</td>
</tr>
<tr>
<td>P12</td>
<td>yecA</td>
<td>tRNA-Leu (taa)</td>
<td>20,043</td>
<td>4,265,444–4,273,677</td>
</tr>
<tr>
<td>P13</td>
<td>yodB</td>
<td>tRNA-Ser (cga)</td>
<td>43,687</td>
<td>5,631,805–5,647,245</td>
</tr>
<tr>
<td>PLE4</td>
<td>cobU</td>
<td>Hypothetical</td>
<td>13,348</td>
<td>6,884,489–6,897,837</td>
</tr>
<tr>
<td>P14 (stx2c)</td>
<td>yeeW</td>
<td>stbB</td>
<td>60,586</td>
<td>3,698,735–3,759,321</td>
</tr>
<tr>
<td>P15a</td>
<td>yehV</td>
<td>mtkA</td>
<td>52,982</td>
<td>3,900,672–3,953,654</td>
</tr>
<tr>
<td>P16 (stx2a)</td>
<td>yfcV</td>
<td>argW</td>
<td>74,662</td>
<td>4,190,350–4,265,012</td>
</tr>
<tr>
<td>P17</td>
<td>argW</td>
<td>lacY</td>
<td>8,233</td>
<td>4,265,444–4,273,677</td>
</tr>
<tr>
<td>P18</td>
<td>ssrA</td>
<td>alpA</td>
<td>20,904</td>
<td>4,549,827–4,570,731</td>
</tr>
<tr>
<td>PLE5</td>
<td>tRNA-Phe (gaa)</td>
<td>pitB</td>
<td>23,331</td>
<td>4,924,934–4,948,266</td>
</tr>
<tr>
<td>P19</td>
<td>yicO</td>
<td>yicC</td>
<td>15,440</td>
<td>5,631,805–5,647,245</td>
</tr>
<tr>
<td>PLE6 (Lee)</td>
<td>selC</td>
<td>yicL</td>
<td>43,021</td>
<td>5,665,860–5,708,881</td>
</tr>
</tbody>
</table>

**Notes:**

- PLE6 (Lee) refers 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).
- PLE7 refers to the end in which the Integrase gene (intA) is located.

---

Frontiers in Microbiology | www.frontiersin.org
October 2020 | Volume 11 | Article 577658
numbers, and NCBI BioProject.

Secondly, Pilon v1.22 (Walker et al., 2014) using the –methylation-aware = dam,dcm and –min-candidate-
5s for each respective sample accounting for methylation
et al., 2015) using both the trimmed ONT FASTQs and
in a three-step process first, using Nanopolish v0.11.1 (Loman
(2017). Polishing of the assembly was performed
Trimmed ONT FASTQ files were assembled using Canu v1.7.1. The

The barcode and y-adapter from each sample’s reads were
generated using Nanoplot v1.8.1 (De Coster et al., 2018).
Run metrics were multiplexed again using Deepbinner 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-adapter 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

Trimmend 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.
FIGURE 2 | Maximum likelihood phylogenetic tree of strains belonging to sub-lineage IIa, most closely related to the outbreak cluster, showing SRA accession, stx type profile, and travel history where available. *denotes Oxford Nanopore Technology sequenced sample 194195.
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 (>10^5 cfu/g), as defined by Regulation (EC) No. 178/2002^1, 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 belonged to a domestic or 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, Ib, and I/IIf 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...
FIGURE 4 | Mid-rooted neighbor joining tree of Shiga toxin-encoding prophages based on Jaccard distance produced from Mash. Strains are annotated as Strain ID, length, and stx type profile. Strains are colored by lineage—green: Ia, red: Ic, blue: I/IIa, gray: I/IIb, orange: IIa, purple: IIb. *denotes Oxford Nanopore Technology sequenced sample 194195.
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 Supplemental Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2020.577658/full#supplementary-material


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.