MICROBIAL GENOMICS

RESEARCH ARTICLE
Dallman *et al.*, *Microbial Genomics* 2021;7:000544
DOI 10.1099/mgen.0.000544

Phylogenetic structure of Shiga toxin-producing *Escherichia coli* O157:H7 from sub-lineage to SNPs

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

Abstract

Sequence similarity of pathogen genomes can infer the relatedness between isolates as the fewer genetic differences identified between pairs of isolates, the less time since divergence from a common ancestor. Clustering based on hierarchical single linkage clustering of pairwise SNP distances has been employed to detect and investigate outbreaks. Here, we evaluated the evidence-base for the interpretation of phylogenetic clusters of Shiga toxin-producing *Escherichia coli* (STEC) O157:H7. Whole genome sequences of 1193 isolates of STEC O157:H7 submitted to Public Health England between July 2015 and December 2016 were mapped to the Sakai reference strain. Hierarchical single linkage clustering was performed on the pairwise SNP difference between all isolates at descending distance thresholds. Cases with known epidemiological links fell within 5-SNP single linkage clusters. Five-SNP single linkage community clusters where an epidemiological link was not identified were more likely to be temporally and/or geographically related than sporadic cases. Ten-SNP single linkage clusters occurred infrequently and were challenging to investigate as cases were few, and temporally and/or geographically dispersed. A single linkage cluster threshold of 5-SNPs has utility for the detection of outbreaks linked to both persistent and point sources. Deeper phylogenetic analysis revealed that the distinction between domestic UK and imported isolates could be inferred at the sub-lineage level. Cases associated with domestically acquired infection that fall within clusters that are predominantly travel associated are likely to be caused by contaminated imported food.

DATA SUMMARY

The sequences for all isolates in this study were stored in the National Center for Biotechnology Information (NCBI) Short Read Archive (SRA) BioProject: PRJNA315192.

INTRODUCTION

Shiga toxin-producing *Escherichia coli* (STEC) O157:H7 emerged as a gastrointestinal pathogen of public health concern in England after being identified as the aetiological agent of a series outbreaks of haemolytic uraemic syndrome (HUS) during the 1980s [1]. HUS is a severe and potentially fatal systemic condition primarily affecting the kidneys [2]. STEC O157:H7 is zoonotic with ruminants, mainly cattle and sheep, being the main animal reservoir [3]. Transmission to humans occurs following the consumption of contaminated food or water, and direct contact with animals or their environment. The infectious dose is low (10–100 bacteria), and person-to-person spread can occur in households and institutional settings.

Whole genome sequencing (WGS) data can be utilized to facilitate public health surveillance of STEC O157:H7 [4]. Sequence similarity of pathogen genomes can infer the relatedness between isolates as the fewer genetic differences identified between pairs of isolates, the less time since divergence from a common ancestor. As such, isolates with very similar genomes have an increased likelihood that they are transmitted via the same vehicle and/or from the same source population [5]. Clustering based on hierarchical single linkage clustering of pairwise SNP distances has been employed to detect and investigate outbreaks [6].

In 2012, 334 STEC O157:H7 isolated from human faecal specimens were sequenced, and each SNP clustering profile was

Received 18 September 2020; Accepted 15 February 2021; Published 15 March 2021
Author affiliations: [1]National Infection Services, Public Health England, 61 Colindale Avenue, London, NW9 5EQ, UK.
***Correspondence:** Claire Jenkins, claire.jenkins1@phe.gov.uk
Keywords: Shiga toxin-producing *Escherichia coli* O157:H7; surveillance; whole genome sequencing.

Abbreviations: GBRU, Gastrointestinal Bacterial Reference Unit; HUS, haemolytic uraemic syndrome; PHE, Public Health England; STEC, Shiga toxin-producing *Escherichia coli*; Stx, Shiga toxin; WGS, whole genome sequencing.

Data statement: All supporting data, code and protocols have been provided within the article or through supplementary data files. One supplementary table is available with the online version of this article.

linked to the available epidemiological data [4]. The pairwise SNP distance distribution revealed that epidemiologically linked isolates, either from the same person, household or outbreaks caused by a known common source, all belonged to 5-SNP single linkage clusters. This definition has been useful for the detection and investigation of outbreaks in the UK both in the presence and absence of epidemiological data from all cases linked to a sequenced isolate of STEC O157:H7 [5]. The aim of this study was to re-evaluate the evidence-base for the interpretation of phylogenetic clusters and strain relatedness at the 5-SNP level, and to explore whether microbiological and epidemiological data linked to deeper phylogenetic relationships can be used to inform public health surveillance.

METHODS

Bacterial strains

There were 1193 isolates of STEC O157:H7 submitted to the Gastrointestinal Bacterial Reference Unit (GBRU), at Public Health England (PHE), for confirmation and typing from 1098 individual patients between July 2015 and December 2016 (Table S1, available in the online version of this article). Strains of STEC O157:H7 were isolated from faecal specimens following culture on cefixime-tellurite sorbitol MacConkey agar (CT-SMAC), from hospital and community cases of gastrointestinal disease submitted to local hospital laboratories in England.

Whole genome sequencing

At GBRU, the cultures were inoculated into nutrient broth and propagated overnight at 37 °C. Genomic DNA from isolates of STEC O157:H7 was extracted on the QiaSymphony system (Qiagen). The sequence library was prepared using the Nextera XT kit and sequenced on the HiSeq 2500 platform (Illumina) yielding paired-end reads of 100 bp in length. High-quality reads were mapped to the reference STEC O157:H7 strain, Sakai (GenBank accession BA000007), using Burrows-Wheeler Aligner – Maximum Exact Matching [BWA MEM (v0.7.2)] [7]. The sequence alignment map outputs from BWA were sorted and indexed to produce a binary alignment map (BAM) using Samtools (v1.1) [8]. Genome Analysis Toolkit (GATK v2.6.5) was then used to create a variant call format (VCF) file from each of the sorted BAMs, which were further parsed to extract only SNP positions of high quality [mapping quality (MQ) >30, depth (DP) >10, variant ratio >0.9] [9]. Any variants called at positions that were within the known prophages in Sakai were masked from further analyses. The remaining variants were imported into SnapperDB v0.2.5 [6]. Hierarchical single linkage clustering was performed on the pairwise SNP difference between all isolates at descending distance thresholds (Δ250, Δ100, Δ50, Δ25, Δ10, Δ5, Δ0) [6]. The result of the clustering is an SNP profile, or SNP address, that is used to describe the population structure based on clonal group membership, as indicated by the number at each level of the seven-number SNP address.

An alignment of polymorphic positions was used to create approximate maximum-likelihood trees using FigTree under the Jukes–Cantor model of nucleotide evolution (http://tree.bio.ed.ac.uk/software/figtree/). Pairwise SNP distances between the genomes of each strain were calculated. Lineage and sub-lineage assignments were performed based on discriminatory SNPs, extracted directly from SnapperDB v0.2.5, that define the population structure, as described previously [4, 6].

Impact Statement

Public Health England operates an enhanced surveillance system for Shiga toxin-producing *Escherichia coli* (STEC) O157:H7 using whole genome sequencing (WGS) data to detect and investigate outbreaks of gastrointestinal disease. We evaluated over 1000 genome sequences of STEC O157:H7 and showed that the majority of isolates belonging to the same patient, same household or same outbreak fell within the same 5-SNP single linkage cluster, and those isolates within these context that fell outside the 5-SNP threshold were due to phage-mediated recombination events. Ten-SNP single linkage clusters were infrequently detected and, with one exception, comprised small numbers of cases. The deeper phylogenetic analysis in this study revealed that the distinction between domestic UK and imported isolates could be inferred at the sub-lineage level. This is particularly informative during foodborne outbreak investigations as these data can provide direction on whether the infected food is domestically produced or imported.

Serotypes were derived from the genome data using the GeneFinder tool, based on the Serotypefinder database [10] and the best match to each of the O and H determinants was reported, as described previously [11]. Sequence type (ST) assignment was performed using the Metric Orientated Sequence Typer (MOST), available from https://github.com/phe-bioinformatics/MOST. Shiga toxin (Stx) subtyping was performed as previously described [12]. Visualization of distributions of pairwise SNP distances and statistical detection of outliers was analysed in Python using plotly (https://plotly.com/) and numpy (https://numpy.org/) respectively. Sunburst representations of SNP clustering were produced using plotly within Python.

The sequences for all isolates in this study were stored in the National Center for Biotechnology Information Short Read Archive BioProject PRJNA248042).

Data collection

Microbiological typing data, including serotype, sequence type and SNP type, and patient demographic data including, sex, age, residential area and recent travel were stored in an in-house integrated molecular national surveillance database. Travel-associated cases were defined as those reporting recent foreign travel to any country outside the UK 7 days prior to

Table 1. Number of isolates from individual patients categorized as sporadic, outbreak and household and 5-SNP cluster within each STEC O157:H7 sub-lineage

Sub-lineage	No. of isolates	Sporadic	Outbreak	Household	5-SNP cluster no known source	No. of clusters at the 250 SNP level
Ia	30	23	0	3	4	9
Ib	25	17	0	4	4	4
Ic	229	101	13	45	72	4
IIa	320	100	157	60	16	28
IIb	140	49	57	24	12	4
IIc	286	178	0	29	78	1
I/II	37	26	0	8	2	3
NSF	7	5	0	2	0	2
SF	23	13	0	7	2	5
Total	1097	512	232	182	190	

NSF – non-sorbitol-fermenting isolates that fall outside the three main lineages.
SF – sorbitol-fermenting isolates that fall outside the three main lineages.

the onset of symptoms, based on information from laboratory reports.

Case definitions

Household: a case who shared the same household as another case.

Outbreak: a case belonging to a cluster of cases where the source of the outbreak was determined.

5-SNP Community Cluster: a case with an isolate belonging to the same 5-SNP single linkage cluster as an isolate from another case where a common source was not determined

Sporadic: a case with an isolate that did not belong to a 5-SNP single linkage cluster as an isolate from another case.

In Table S1 isolates belonging to the same case or the same household were assigned the same number in the columns headed ‘Same patient’ and ‘Same household’, respectively. Isolates linked to an outbreak where the source was determined were labelled A–F in the column labelled ‘Outbreaks’. The letters correspond with the outbreaks described in the text below.

RESULTS

Overview of the phylogenetic relatedness of isolates of STEC O157:H7 at the sub-lineage level

Of the 1097 isolates from individual patients in this study, 512/1097 (46.7%) were identified as belonging to a sporadic case, 182/1097 (16.6%) were from cases belonging to household clusters, 232/1097 (21.1%) were linked to an outbreak where the source of the outbreak was determined, and 190/1097 (17.4%) fell within a 5-SNP single linkage cluster of another isolate where a common source was not determined. Of the 512 sporadic cases, 494/512 (96.5%) belonged to one of the seven established STEC O157:H7 sub-lineages (Ia-Ic, IIa-IIc, I/II) and 18/512 (3.5%) fell outside of this classification structure (Table 1). The distribution of 5-SNP single linkage clusters in each sub-lineage at six different levels of relatedness at the 250-, 100-, 50-, 25-, 10- and 5-SNP levels [6], is shown in Fig. 1. The majority of sporadic isolates belonged to sub-lineage IIc (178/512, 34.8%), whereas the majority of isolates that fell within a 5-SNP cluster, including those that were linked to an outbreak or household, belonged to IIa (233/610, 38.2%) (Table 1).

The distribution of 5-SNP single linkage clusters in each sub-lineage is shown again in Fig. 2, with each 5-SNP single linkage cluster represented by a single isolate. In this figure, the segments are coloured based on the proportion of cases reporting foreign travel within 7 days of the onset of symptoms. The majority of clusters that belong to sub-lineages Ic, IIb and I/II are domestically acquired (i.e. acquired in the UK). In contrast, sub-lineage Ia and Ib are almost exclusively associated with cases reporting recent travel outside the UK, who were likely to have acquired their infection abroad.

Sub-lineages IIa and IIc display a mixture of domestically acquired and travel-associated clusters. However, within each sub-lineage, at more discriminatory SNP levels, clear associations to either travel or domestic acquisition can be identified. For example, in lineage IIa there is a 50-SNP-level cluster (designated 5.156.490. in Fig. 2) that is almost exclusively domestically acquired, whilst the majority of the remaining clusters within sub-lineage IIa are travel-associated. Cases associated with domestically acquired infection that fall within clusters that are

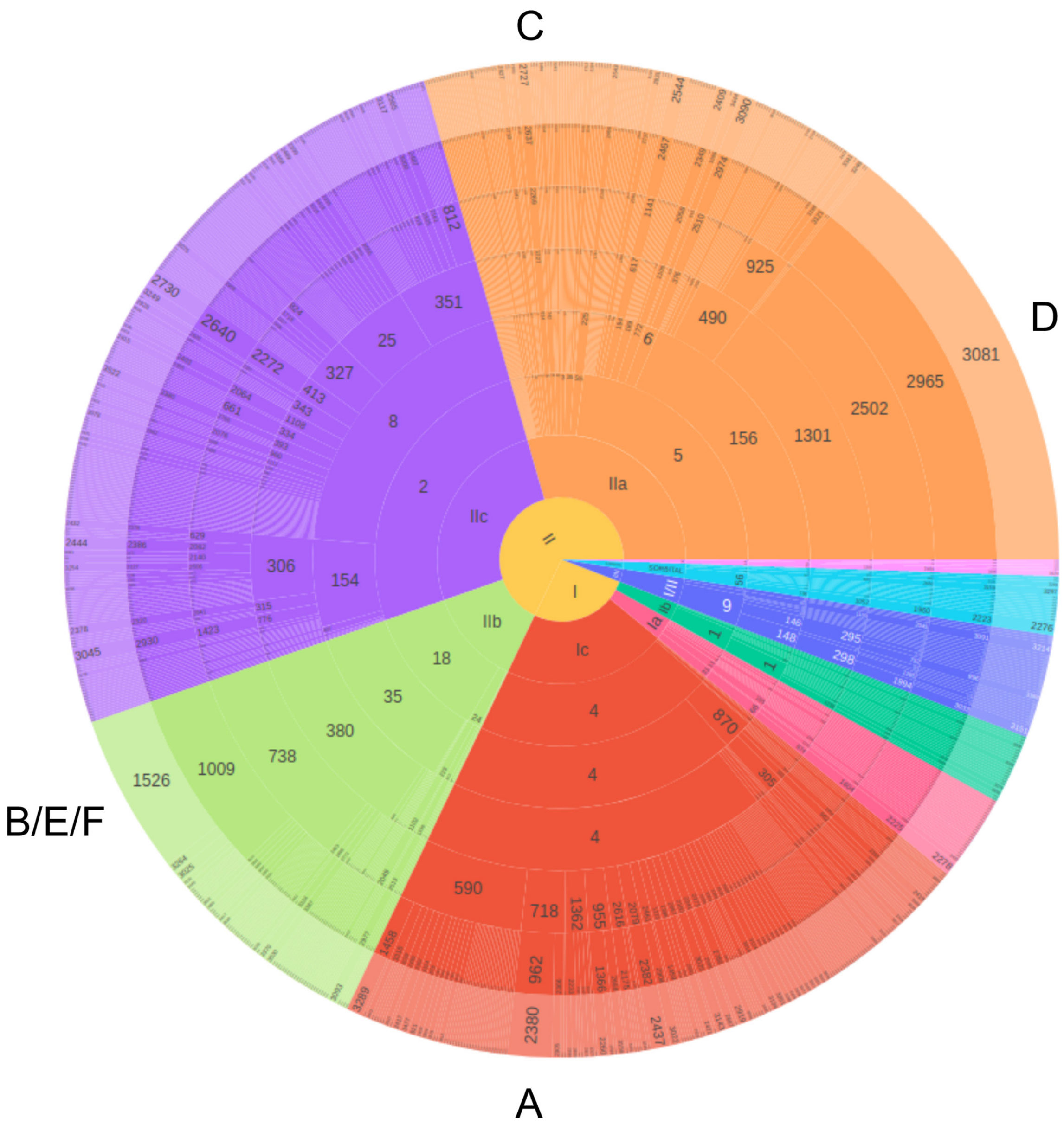

Fig. 1. Sunburst diagram showing the distribution of isolates belonging to a lineage and sub-lineage, and six descending concentric circles represent single linkage SNP clusters at the 250-, 100-, 50-, 25-, 10- and 5-SNP levels. The size of each segment represents the proportion of isolates in that cluster. The numbers represent the single linkage SNP cluster designation at each level. For example, the SNP type or SNP address for outbreak D is: 5.156.1301.2502.2986.3081 [6]. Segments are coloured based on sub-lineage. Segments representing the outbreaks are labelled A–F and described in detail in Table 2.

predominantly travel-associated are likely to be caused by contaminated imported food. Similarly, sub-lineage IIc comprises both domestically acquired and travel-associated isolates (Fig. 2). However, these two groups can be differentiated at more discriminatory levels of SNP clustering. For example, at the 100-SNP level, most isolates belonging to 2.154. are travel-associated, whereas for isolates belonging to 2.8., the travel-association is revealed at the 50-SNP level in 2.8.351. (Fig. 2).

Phylogenetic context of outbreaks of STEC O157:H7

There were 232 isolates that were part of six outbreaks for which an epidemiological link between cases was ascertained. These included outbreaks linked to (A) cross-contamination of raw and cooked meat at a butcher's shop [13], (B) contaminated pre-packed salad sold by a national retailer [14], (C) contaminated frozen, grated coconut [15], (D) imported Red Batavia leaves [16], (E) undercooked lamb meat sausages at a restaurant [14] and

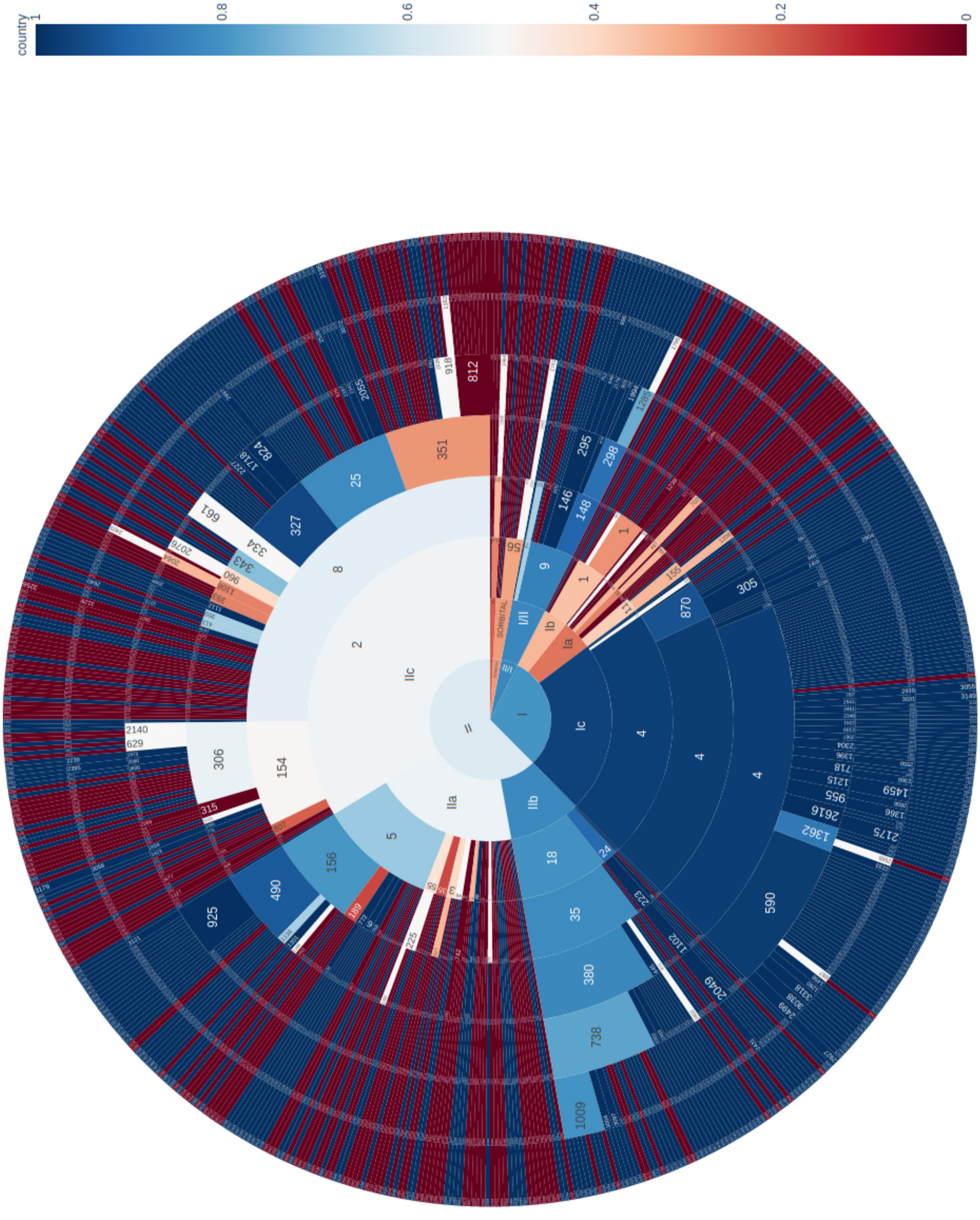

Fig. 2. Sunburst diagram showing the distribution of isolates belonging to a lineage and sub-lineage, and six descending concentric circles represent single linkage SNP clusters at the 250-, 100-, 50-, 25-, 10- and 5-SNP levels. The segments were coloured based on the proportion of isolates from cases reporting foreign travel with 7 days of the onset of symptoms, with dark blue being no cases reporting recent travel and dark red being all cases reporting recent travel outside the UK. 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.

Table 2. Characteristics of documented outbreaks of STEC O157:H7 linked to known source July 2015 to December 2016

Outbreak	Description (phage type)	No. of cases	Median time between receipt of isolates (days)	Pairwise SNP distances			Time range (days)	Average case dispersal (km)
				Median	Minimum	Maximum		
A	Cross-contamination of raw and cook meat at a butcher's shop in the North East (PT21/28)	13	1	1	0	4	21	19
B	National foodborne outbreak caused by contaminated pre-packed salad (PT8)	45	0	0	0	3	55	192
C	Contaminated coconut (PT24)	5	1	1	0	2	8	131
D	Italian Red Batavia (PT34)	157	1	1	0	25	117	167
E	Contaminated lamb meat sausages (PT8)	4	2	1.5	0	3	82	204
F	Contaminated lamb mince (PT8)	8	4	4	0	12	224	170

(F) contaminated lamb mince sold by a national retailer [14] (Table 2).

The majority of clusters that belong to sub-lineages Ic, IIb and I/II are domestically acquired, and this is consistent with the known sources of the four outbreaks caused by strains belonging to lineages Ic (outbreak A, UK beef cattle) and IIb (outbreaks B, E and F, domestically produced salad vegetables and UK lamb). Outbreak D was caused by contaminated imported salad leaves and this is consistent with its non-domestic context within sub-lineage IIa. Isolates from outbreak D (*n*=159) belonged to sub-lineage IIa and represented the majority of cases in this sub-lineage (157/320, 49.1%). The highest level of sub-lineage diversity, as measured by the number of different 250-SNP single linkage clusters within each sub-lineage, was observed within sub-lineage IIa (Table 1).

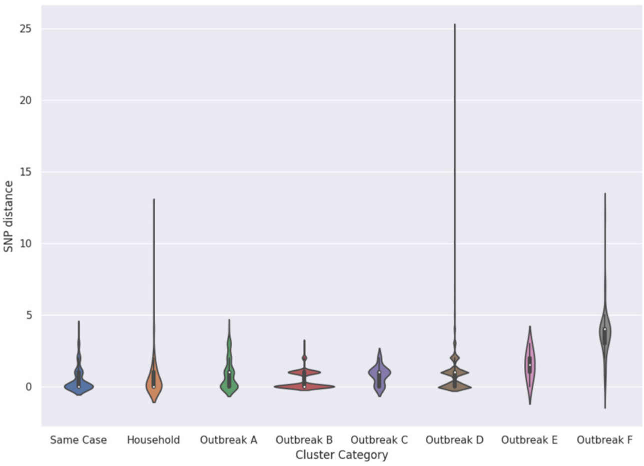

Fig. 3. Violin plot showing the distribution of pair-wise SNP distances from isolates from the same case, isolates from the same household and isolates from outbreaks where an epidemiological link was made.

Distribution of pairwise distance between isolates from outbreaks where the vehicle and/or source of infection was known

Analysing travel data at the sub-lineage level was useful for predicting whether a strain was imported or domestic in origin. However, outbreak detection and case ascertainment during an outbreak investigation requires analysis at a higher level of discrimination. The median pairwise SNP distances between isolates for each outbreak varied between 0 SNPs for outbreak B (minimum SNPs=0, maximum SNPs=3) (pre-packed salad), and 4 SNPs for outbreak F (minimum SNPs=0, maximum SNPs=12) (lamb mince) (Table 2 and Fig. 3). For outbreak D the median pairwise SNP distance was 1 SNP (minimum SNPs=0, maximum SNPs=25). Outbreak D (imported Red Batavia leaves) contained an isolate with an average pairwise SNP distance significantly greater than the other isolates in the outbreak [Z-score 7.82 (*P*<0.05)] (Table 1, Fig. 3). Further analysis of the location of variants in the chromosome revealed that one of the isolates had acquired an additional 24 SNPs mediated by a recombination event in Sakai Phage (SP)7.

Outbreak E (lamb meat sausages) fell within the same 5-SNP single linkage cluster as outbreak B (pre-packed salad), and outbreak F (lamb mince) fell within the same 10-SNP single linkage cluster as outbreaks B and E (Figs 1 and 4). Outbreaks E and F were both caused by the consumption of contaminated lamb meat, most likely from the same source [14]. Outbreak B was caused by the consumption of contaminated salad leaves. Subsequent investigations, including analysis of animal movement data, indicated that the salad may have be cultivated in an environment contaminated by sheep excrement, most likely contaminated irrigation water drawn from a non-potable source [14].

The median time between receipt of isolates varied between 0 days for outbreak B (pre-packed salad) and 23 days for outbreak F (lamb mince) (Table 2). The outbreak with the shortest duration of time between submission of the first and last isolate was outbreak C (frozen, grated coconut) (8 days)

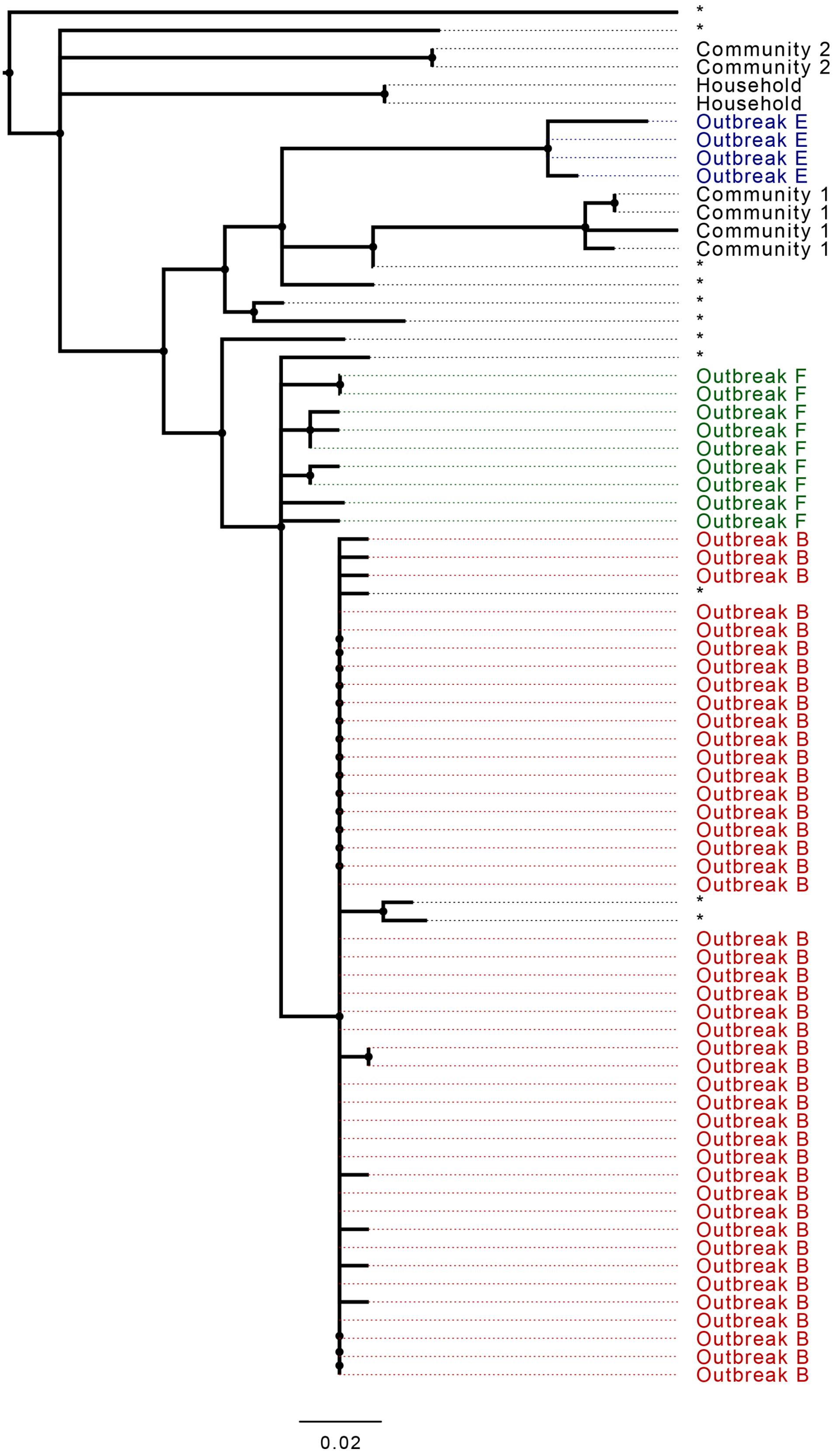

Fig. 4. Maximum-likelihood phylogeny of the 10-SNP single linkage cluster t10:1051. Sporadic cases are labelled*.

whereas outbreak F (lamb mince) had the longest time frame (224 days) (Table 2, Fig. 3). The two salad outbreaks B and D involving the largest numbers of cases were the most temporarily and spatially dispersed, with cases occurring for 69 and 117 days and lived an average of 167 and 197 km from each other, respectively. The durations of the outbreaks caused by contaminated coconut and meat products were 8 and 21 days and the cases lived an average of 131 and 19 km from each other, respectively.

Distribution of pairwise distance between isolates from travel clusters

There were 16 clusters comprising 41 isolates from cases reporting recent travel to the same country (Turkey *n*=5, Spain *n*=2, Egypt *n*=2, Romania *n*=1, United Arab Emirates *n*=1, Ireland *n*=1, Europe other than UK *n*=1, Greece *n*=1, Cyprus *n*=1). For travel-related clusters the median cluster size was two cases (with a maximum of six cases). The median SNP distance between isolates in travel-related cases was 0 SNPs (with a maximum of 5 SNPs) and the median time interval between cases was 7 days.

Analysis of isolates with sequences within the same 5-SNP single linkage cluster not linked to a known outbreak

After removing duplicate isolates from the same patient and selecting the earliest representative from the same household, there were 216 isolates belonging to 64 5-SNP single linkage community clusters that could not be explained by cases reporting the same travel destination or an outbreak linked to a common source. For community clusters, the median cluster size was two cases (with a maximum of six cases). The median SNP distance between isolates in community clusters was 1 SNP (with a maximum of 10 SNPs) and the median time interval between cases was 7 days.

A comparison was made between the geographical spread of each cluster (determined using the average residential distance between cases) and the duration, for both investigated outbreaks and community clusters. The majority of 5-SNP community clusters were short-lived (<50 days) and geographically restricted (<100 km average residential distance). Community clusters with a longer duration (>50 days) were more likely to be geographically dispersed (>100 km average residential distance).

Analysis of isolates with sequences within the same 10-SNP single linkage cluster

After removing duplicate isolates from the same patient and selecting the earliest representative from the same household, there were 112 isolates that belonged to 18 10-SNP single linkage clusters.

The largest 10-SNP single linkage cluster consisted of 75 cases with maximum SNP distance of 37 and persisted for 495 days in the study. Nested within this cluster were five 5-SNP single linkage clusters including three of the outbreaks described above, B, E and F, two community clusters and five household clusters (four within outbreak B) (Fig. 4). In addition to the 52 cases linked to outbreaks B (*n*=45) and E (*n*=7), there were four cases that fell within the same 5-SNP single linkage cluster, three of which were located within outbreak B. However, they were temporally distinct from the original outbreak on September 2015, with all five cases reporting onsets dates in 2016, and none was linked to the consumption of salad (Fig. 4). Previous analysis of this 10-SNP cluster identified multiple transmission routes involving exposures to an ovine source, either consumption of contaminated lamb-based meat products, consumption of salad vegetables cultivated in an environment contaminated with sheep excrement, and/or direct contact with sheep or their environment [14].

For the remaining 17 10-SNP single linkage clusters in the dataset, the median cluster size was two cases (with a maximum of seven cases). The median SNP distance between isolates was 10 SNPs (minimum SNPs=0, maximum SNPs=18), and the median time interval between cases was 106 days.

Distribution of pairwise distance between isolates from the same case or same household

There were 178 isolates from 82 cases where multiple isolates were sequenced from the same person. Of these, 72/82 (87.8%) cases had two isolates, 8/82 (9.8%) cases were linked to three isolates and a further two cases had four and six isolates linked to each. The median time between receipt of serial isolates was 7 days, with a minimum of 0 days and a maximum of 77 days. The median SNP distance between isolates from the same case was 0 SNPs with a maximum of 4 SNPs (Fig. 3).

There were 182 isolates that were part of 80 separate household clusters: 64/80 (80.8%) households had two cases, 14/80 (16.7%) households had three cases and two households had four cases. The median time between receipt of isolates from the same household was 6 days with a minimum of 0 days and a maximum of 29 days. The median SNP distance between isolates from the household was 0 SNPs with a maximum of 12 SNPs (Fig. 3). The 12 SNPs identified between isolates in one household cluster was mediated by three phage-mediated recombination events (3 SNPs in SP9, 7 SNPs in SP11 and 2 SNPs in SP12). If the 12-SNP isolate pair was removed from the analysis, the maximum difference was 4 SNPs.

DISCUSSION

As previously reported, isolates from cases with known epidemiological links, specifically those from the same patient, same household or same outbreak with an established source, for the most part fell within 5-SNP single linkage clusters [4]. Other studies evaluating the utility of SNP clustering to detect outbreaks report similar findings [17–26]. Further analysis revealed that the SNP difference of the isolate pairs from one household and from outbreak D that fell outside the 5-SNP threshold were due to phage-mediated recombination events. STEC O157:H7 has an extensive prophage repertoire comprising up to 15% of the genome [27] with such loci known to be subject to intra- and inter-strain recombination events [28]. As such, every effort must be taken to detect and mask these regions of relatedness during the analysis, as the incorporation of exogenous DNA may distort interpretations of genetic similarity [29].

In this study the term outbreak was used to describe a 5-SNP single linkage cluster where a common vehicle, or exposure, was identified. Each outbreak exhibited different characteristics and included point source contamination events: for example, cross-contamination of pre-cooked meat at a butcher's shop or contamination of salad vegetables due to the use of contaminated irrigation water or flooding, and persistent transmission caused by contaminated meat in the food chain causing sporadic infections in humans due to inadequate cooking either in restaurants or in the home. Outbreaks A, D and E were initially identified by epidemiological links, whereas outbreaks B, C and F were initially detected by routine surveillance of isolates belonging to the same 5-SNP single linage cluster. Not every case linked to the six outbreaks by microbiological typing reported exposure to the implicated vehicle [13, 15, 16]. Explanations for this included poor patient recall, the food item being obscure (such as a minor ingredient or side dish to the main meal [30]), secondary person to person transmission, or an alternative vehicle. However, regardless of the vehicle and the caveats, the 5-SNP threshold was useful for both outbreak detection and case ascertainment.

The vehicle and/or source of the majority of 5-SNP single linkage community clusters was not identified. These unresolved community clusters were more likely to be temporally and/or geographically related than sporadic cases, providing circumstantial evidence that they comprise cases that may have a common exposure. For example, temporal links may conform to a point source contamination event, and spatial links may indicate a common environmental exposure. However, the epidemiological link may be confounded by lack of data on recent food histories and/or animal or environmental exposures, due to the small number of cases associated with each cluster.

In this dataset, 10-SNP single linkage clusters occurred infrequently. With the notable exception of one large cluster which has been described previously [14], 10-SNP single linkage clusters comprised small numbers of cases. For a zoonotic pathogen such as STEC O157:H7, the mode of transmission is predominantly foodborne, or contact with animals and/or their environment. Due to the large unsampled diversity in the animal reservoir, and often transient food-chain contamination, the diversity sampled in human STEC cases is generally well separated and therefore discrete genetic clustering is observed. This is in contrast to the clustering observed in human clinical isolates of non-typhoidal *Salmonella*, where persistent clusters at low diversity are seen. These persistent clusters are often, but not always, linked to the poultry industry [31, 32]. Closely related isolates persisting over a prolonged time frame are also observed in human host-restricted gastrointestinal pathogens predominantly transmitted person-to-person via a faecal–oral route, notably the typhoidal Salmonellae, *Shigella flexneri* and *Shigella sonnei* [33].

Although infrequent, 10-SNP single linkage clusters were detected in this dataset. In this study, the large 10-SNP single linkage cluster was probably caused by an endemic domestic source that gave rise to multiple point source outbreaks caused by different food vehicles and reflecting different transmission events over a time frame of 12–18 months. However, without nested point source outbreaks identifying probable food vehicles or geographical links, 10-SNP single linkage clusters may be challenging to investigate, as they are likely to be temporally and often spatially dispersed.

STEC O157:H7 isolated from cases of travellers' diarrhoea are useful in contributing to the evidence-base for identifying non-domestic clusters [5, 16]. This is particularly informative during foodborne outbreak investigations as these data can provide direction on whether the infected food is domestically produced or imported. The deeper phylogenetic analysis in this study revealed that the distinction between domestic UK and imported isolates could be inferred at the sub-lineage level for lineages Ic, I/II and IIb (domestic UK) and Ia, Ib and IIa (imported). Previous studies have shown that the domestic UK lineages have been endemic in UK cattle (Ic and I/II) and sheep (IIb) for decades [34, 35]. For sub-lineage IIc, a higher level of discrimination is required to infer a domestic or imported source, and this may indicate that sub-lineage IIc was more recently imported into the UK [34]. Nevertheless, it is evident that certain clades of sub-lineage IIc are now endemic in UK cattle [34].

SNP typing has proven capability for public health surveillance of zoonotic, foodborne pathogens, such as STEC O157:H7. Human cases appear as discrete clusters within the wider unsampled diversity present in the animal reservoir, and deeper phylogenetic relationships linked to epidemiological data may support hypotheses on the origin of strains causing human disease. A single linkage cluster threshold of 5 SNPs has utility for the detection and investigation of outbreaks linked to both persistent and point sources. However, the epidemiological link may be obscured by lack of data due to poor patient recall or small numbers of cases. Ten-SNP single linkage clusters are more challenging to investigate, as epidemiological links may also be confounded by the cases being more widely dispersed temporally and spatially. Associations with travel data can be used to infer whether a strain is more likely to be domestic or imported, thus providing clues to the likely geographical origin of the strain, and the possible vehicle of infection, during outbreak investigations.

Funding information

The research was part funded by the National Institute for Health Research Health Protection Research Unit in Gastrointestinal Infections at the 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 PHE. The views expressed are those of the authors and not necessarily those of the National Health Service, the NIHR, the Department of Health or PHE.

Author contributions

T.J.D. and C.J. conceptualized the project. T.J.D. and 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.

Conflicts of interest

The authors declare that there are no conflicts of interest.

References

1. Taylor CM, White RH, Winterborn MH, Rowe B. Haemolytic uraemic syndrome: clinical experience of an outbreak in the West Midlands. *Br Med J (Clin Res Ed)* 1986;292:1513–1516
2. Tarr PI, Gordon CA, Chandler WL. Shiga-toxin-producing *Escherichia coli* and haemolytic uraemic syndrome. *Lancet* 2005;365:1073–1086.
3. Byrne L, Jenkins C, Launders N, Elson R, Adak GK. The epidemiology, microbiology and clinical impact of Shiga toxin-producing *Escherichia coli* in England, 2009–2012. *Epidemiol Infect* 2015;143:3475–3487
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. 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.
6. 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.
7. Li H, Durbin R. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* 2010;26:589–595
8. 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
9. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 2014;30:1312–1313
10. Joensen KG, Tetzschner AM, Iguchi A, Aarestrup FM, Scheutz F. Rapid and easy *In Silico* Serotyping of *Escherichia coli* isolates by use of Whole-Genome sequencing data. *J Clin Microbiol* 2015;53:2410–2426.
11. Chattaway MA, Dallman TJ, Gentle A, Wright MJ, Long SE *et al*. Whole genome sequencing for public health surveillance of Shiga toxin-producing *Escherichia coli* other than serogroup O157. *Front Microbiol* 2016;7:258.
12. Ashton PM, Perry N, Ellis R, Petrovska L, Wain J *et al*. Insight into Shiga toxin genes encoded by *Escherichia coli* O157 from whole genome sequencing. *PeerJ* 2015;3:e739.
13. Wilson D, Dolan G, Aird H, Sorrell S, Dallman TJ *et al*. Farm-to-fork investigation of an outbreak of Shiga toxin-producing Escherichia coli O157. *Microb Genom* 2018;4:e000160.
14. 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
15. Greig DR, Mikhail AFW, Dallman TJ, Jenkins C. An outbreak of a rare strain of Shiga toxin-producing *Escherichia coli* O157:H7 *stx2a/stx2c* linked to grated frozen coconut. *Front Microbiol*.
16. 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;23:17–197.
17. Sadiq SM, Hazen TH, Rasko DA, Eppinger M. EHEC genomics: past, present, and future. *Microbiol Spectr* 2014;2:0020–2013.
18. Joensen KG, Scheutz F, Lund O, Hasman H, Kaas RS *et al*. Real-time whole-genome sequencing for routine typing, surveillance, and outbreak detection of verotoxigenic *Escherichia coli*. *J Clin Microbiol* 2014;52:1501–1510.
19. Holmes A, Allison L, Ward M, Dallman TJ, Clark R *et al*. Utility of whole-genome sequencing of *Escherichia coli* O157 for outbreak detection and epidemiological surveillance. *J Clin Microbiol* 2015;53:3565–3573.
20. Ferdous M, Friedrich AW, Grundmann H, de Boer RF, Croughs PD *et al*. Molecular characterization and phylogeny of Shiga toxin-producing *Escherichia coli* isolates obtained from two Dutch regions using whole genome sequencing. *Clin Microbiol Infect* 2016;22:642.e1–64642.
21. Parsons BD, Zelyas N, Berenger BM, Chui L. Detection, characterization, and typing of Shiga toxin-producing *Escherichia coli*. *Front Microbiol* 2016;7.
22. Rusconi B, Sanjar F, Koenig SS, Mammel MK, Tarr PI *et al*. Whole Genome sequencing for Genomics-Guided investigations of *Escherichia coli* O157:H7 outbreaks. *Front Microbiol* 2016;7:985.
23. Gilchrist CA, Turner SD, Riley MF, Petri WA, Hewlett EL. Whole-Genome sequencing in outbreak analysis. *Clin Microbiol Rev* 2015;28:541–563.
24. Rowell S, King C, Jenkins C, Dallman TJ, Decraene V *et al*. An outbreak of Shiga toxin-producing *Escherichia coli* serogroup O157 linked to a lamb-feeding event. *Epidemiol Infect* 2016;144:2494–2500.
25. 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.
26. 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.
27. Hayashi T, Makino K, Ohnishi M, Kurokawa K, Ishii K. Complete genome sequence of enterohemorrhagic *Escherichia coli* O157:H7 and genomic comparison with a laboratory strain K-12. *DNA Res* 2001;8:11–22.
28. 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.
29. Collins C, Didelot X. Reconstructing the ancestral relationships between bacterial pathogen genomes. *Methods Mol Biol* 2017;1535:109–137.
30. Byrne L, Adams N, Glen K, Dallman TJ, Kar-Purkayastha I *et al*. Epidemiological and microbiological investigation of an outbreak of severe disease from Shiga toxin–producing *Escherichia coli* O157 infection associated with consumption of a Slaw Garnish. *J Food Prot* 2016;79:1161–1168.
31. Pijnacker R, Dallman TJ, Tijsma ASL, Hawkins G, Larkin L *et al*. An international outbreak of *Salmonella enterica* serotype enteritidis linked to eggs from Poland: a microbiological and epidemiological study. *Lancet Infect Dis* 2019;19:778–786.
32. Dallman T, Inns T, Jombart T, Ashton P, Loman N *et al*. Phylogenetic structure of European *Salmonella* Enteritidis outbreak correlates with national and international egg distribution network. *Microb Genom* 2016;2:e000070.
33. Bardsley M, Jenkins C, Mitchell HD, Mikhail AFW, Baker KS *et al*. Persistent transmission of *Shigellosis* in England is associated with a recently emerged multidrug-resistant strain of *Shigella sonnei*. *J Clin Microbiol* 2020;58:pii: e01692–19.
34. 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.
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.