JCM
Journals.ASM.org

Utility of Whole-Genome Sequencing of *Escherichia coli* O157 for Outbreak Detection and Epidemiological Surveillance

Anne Holmes,[a] Lesley Allison,[a] Melissa Ward,[c] Timothy J. Dallman,[d] Richard Clark,[b] Angie Fawkes,[b] Lee Murphy,[b] Mary Hanson[a]

Scottish *E. coli* O157/VTEC Reference Laboratory (SERL), Royal Infirmary of Edinburgh, Edinburgh, Scotland[a]; Wellcome Trust Clinical Research Facility (WTCRF), University of Edinburgh, Western General Hospital, Edinburgh, Scotland[b]; Centre for Immunity, Infection and Evolution, School of Biological Sciences, University of Edinburgh, Edinburgh, Scotland[c]; Gastrointestinal Bacterial Reference Unit, Microbial Services Division, Public Health England, London, United Kingdom[d]

Detailed laboratory characterization of *Escherichia coli* O157 is essential to inform epidemiological investigations. This study assessed the utility of whole-genome sequencing (WGS) for outbreak detection and epidemiological surveillance of *E. coli* O157, and the data were used to identify discernible associations between genotypes and clinical outcomes. One hundred five *E. coli* O157 strains isolated over a 5-year period from human fecal samples in Lothian, Scotland, were sequenced with the Ion Torrent Personal Genome Machine. A total of 8,721 variable sites in the core genome were identified among the 105 isolates; 47% of the single nucleotide polymorphisms (SNPs) were attributable to six "atypical" *E. coli* O157 strains and included recombinant regions. Phylogenetic analyses showed that WGS correlated well with the epidemiological data. Epidemiological links existed between cases whose isolates differed by three or fewer SNPs. WGS also correlated well with multilocus variable-number tandem repeat analysis (MLVA) typing data, with only three discordant results observed, all among isolates from cases not known to be epidemiologically related. WGS produced a better-supported, higher-resolution phylogeny than MLVA, confirming that the method is more suitable for epidemiological surveillance of *E. coli* O157. A combination of *in silico* analyses (VirulenceFinder, ResFinder, and local BLAST searches) were used to determine *stx* subtypes, multilocus sequence types (15 loci), and the presence of virulence and acquired antimicrobial resistance genes. There was a high level of correlation between the WGS data and our routine typing methods, although some discordant results were observed, mostly related to the limitation of short sequence read assembly. The data were used to identify sublineages and clades of *E. coli* O157, and when they were correlated with the clinical outcome data, they showed that one clade, Ic3, was significantly associated with severe disease. Together, the results show that WGS data can provide higher resolution of the relationships between *E. coli* O157 isolates than that provided by MLVA. The method has the potential to streamline the laboratory workflow and provide detailed information for the clinical management of patients and public health interventions.

Shiga toxin-producing *Escherichia coli* (STEC) strains are important gastrointestinal pathogens and common causes of acute renal failure in children worldwide (1, 2, 3). In Scotland, the epidemiology and clinical outcome of *E. coli* O157 infection in humans has been closely monitored by Health Protection Scotland (HPS) since the introduction of enhanced surveillance in 1999, following a rapid increase in the number of microbiologically confirmed cases of STEC infection during the mid-1990s. Enhanced surveillance was extended in 2003 to include non-O157 STEC, and enhanced surveillance of hemolytic-uremic syndrome (HUS) was established in 2003. The most common STEC serogroup isolated from patients in Scotland is *E. coli* O157. In the past 5 years, the number of culture-confirmed cases of *E. coli* O157 infection in Scotland ranged from 195 to 263 per year, compared with 22 to 75 cases of non-O157 STEC infection per year (4, 5). On average, 43% of the STEC infection cases in Scotland require hospitalization and 9% progress to HUS, mostly in children (5). The reported rate of *E. coli* O157 infection in Scotland (4.9 cases per 100,000 population in 2014 [HPS personal communication]) is higher than in most countries, including other countries within the United Kingdom (5, 6). The reasons for this are unclear and likely to be multifactorial but may be related to the relative densities of human and cattle populations, which are the main reservoir of *E. coli* O157. Although large outbreaks of infections associated with food products have been reported in Scotland (7, 8, 9), the majority of infections are apparently sporadic, and contact with animal feces, exposure to farm animals or farm environments, and drinking of water from private water supplies have all been identified as strong risk factors for infection (10, 11).

The continued use of appropriate control measures, which is dependent on the prompt diagnosis of cases and detection of outbreaks, is essential to limit the spread of this pathogen. The Scottish *Escherichia coli* O157/Verotoxigenic *E. coli* Reference Laboratory (SERL) works closely with Scottish National Health Service (NHS) boards and with HPS and plays a pivotal role in case confirmation and outbreak detection by providing identification and typing services for *E. coli* O157 and other STEC isolates. Putative isolates of *E. coli* O157 are referred to the SERL from all 14 Scottish health boards for confirmation of identity and typing. In addition,

Received 4 May 2015 **Returned for modification** 13 June 2015
Accepted 30 August 2015

Accepted manuscript posted online 9 September 2015

Citation Holmes A, Allison L, Ward M, Dallman TJ, Clark R, Fawkes A, Murphy L, Hanson M. 2015. Utility of whole-genome sequencing of *Escherichia coli* O157 for outbreak detection and epidemiological surveillance. J Clin Microbiol 53:3565–3573. doi:10.1128/JCM.01066-15.

Editor: D. J. Diekema

Address correspondence to Anne Holmes, anne.holmes@luht.scot.nhs.uk.

Supplemental material for this article may be found at http://dx.doi.org/10.1128/JCM.01066-15.

fecal samples from cases with typical symptoms of STEC, but culture negative in local diagnostic laboratories, are sent to the SERL for further analysis. Currently the SERL uses a range of phenotypic and genotypic typing techniques, including real-time-PCR, phage typing, disc diffusion susceptibility testing, multilocus variable-number tandem-repeat analysis (MLVA), and pulsed-field gel electrophoresis (PFGE). However, there is an ever-increasing body of evidence demonstrating the potential benefits of WGS as a tool that may replace these methods (12, 13, 14, 15, 16). Bacterial WGS is becoming increasingly affordable and has demonstrated better resolution than more conventional typing methods that sample only a fraction of the genome (17, 18, 19). As a reference laboratory, it is essential that we continually aim to improve our services to provide precise and timely data to support public health interventions. In this study, we have compared our current methods with WGS for the typing and characterization of *E. coli* O157 for epidemiological surveillance and detection of outbreaks. In addition, we have correlated genome content with pathogenicity and ancestry to try to achieve a better understanding of the genetic factors associated with virulence in STEC.

MATERIALS AND METHODS

Bacterial isolates. Isolates of *E. coli* O157 ($n = 105$) from human fecal samples received in Lothian laboratories between 1 April 2007 and 31 March 2012 were analyzed in this study (see Table S1 in the supplemental material). These included isolates from 10 cases received over an 11-month period in 2011 associated with a United Kingdom-wide outbreak that was linked to unwashed vegetables (20), isolates that were epidemiologically linked by place and time (eight single-household clusters, one cluster involving two households that was farm related, and one travel-associated cluster), and sporadic isolates from 27 patients thought to have acquired their infections outside the United Kingdom. The patients had all been interviewed by using standardized questionnaires to collect information, including severity of infection and travel history, to identify any epidemiological links and try to pinpoint the source of infection. The ages of the patients sampled ranged from 1 to 85 years, with a gender distribution of 42% male and 58% female.

DNA isolation. *E. coli* O157 isolates were incubated overnight at 37°C on sorbitol MacConkey agar (Oxoid Ltd., Basingstoke, United Kingdom). DNA was extracted with the Wizard Genomic DNA purification kit (Promega Ltd. UK, Southampton, United Kingdom) as described by the manufacturer. The quality of the genomic DNA (gDNA) was checked by gel electrophoresis, and the quantity and purity were measured with the NanoDrop 1000 apparatus (NanoDrop Products, Thermo Scientific).

Phenotypic testing. Phage typing was performed with 16 phages as previously described (21). β-Glucuronidase (β-GUD) production was assessed by using TBX agar (tryptone bile X-glucuronide agar; E&O Laboratories); ATCC 25922 was used as a control strain. Antibiotic sensitivity patterns were determined by the disc diffusion method with 15 antibiotics routinely tested in the SERL for surveillance purposes: chloramphenicol, ciprofloxacin, ampicillin, gentamicin, streptomycin, meropenem, nalidixic acid, kanamycin, tetracycline, trimethoprim, piperacillin-tazobactam, cefotaxime, ceftazidime, co-amoxiclav and co-trimoxazole. The European Committee on Antibiotic Susceptibility Testing (EUCAST) criteria were used to determine resistance.

MLVA. MLVA was performed as previously described (22). Raw data (.fsa files) from an ABI 3130 genetic analyzer (Applied Biosystems) were imported and analyzed in BioNumerics v6.6 (Applied Maths, Sint-Martens-Latem, Belgium) with the MLVA plugin. A minimum spanning tree was produced with the MLVA allele numbers.

Shiga toxin gene subtyping. PCR assays and gel electrophoresis were performed as described by Scheutz et al. (23), with a few modifications. A sample volume of 15 μl with a HotStarTaq master mix kit (Qiagen UK Ltd., Crawley, United Kingdom), 0.2 μM primer(s), and 1.5 μl of DNA was used. stx_{1a}, stx_{1c}, and stx_{1d} primers were multiplexed in a single reaction mixture. stx_2 primers for stx_{2b}, stx_{2e}, and stx_{2g} were multiplexed, while stx_{2a}, stx_{2c}, stx_{2d}, and stx_{2f} were used in singleplex reaction mixtures. The thermocycler conditions were 95°C for 15 min and then 35 cycles of 94°C for 50 s, 66°C for 40 s, and 72°C for 3 min. Amplicons were run on a 2% agarose gel.

WGS. WGS was performed with the Ion Torrent Personal Genome Machine (PGM; Life Technologies, Carlsbad, CA) at the Wellcome Trust Clinical Research Facility, Edinburgh, Scotland. Samples were sequenced retrospectively over a period of ~1 year on the same machine by the same personnel and analyzed with the same version of the software to limit the effects of batching. Libraries were generated with 1 μg of the gDNA and enzyme fragmented with the Ion XpressPlus Fragment Library kit. Specifically, gDNA was fragmented into 200- to 300-bp blunt-ended DNA fragments. The fragmented DNA was ligated to Ion Torrent-compatible barcoded adapters; this was followed by nick repair to complete the linkage between the adapters and DNA inserts. The adapter-ligated library was then size selected for optimum length (330 bp), and the final library was amplified. An aliquot of the library was analyzed on the Bioanalyzer instrument with an Agilent High Sensitivity DNA kit to assess the size distribution and determine the molar library concentration. Two barcoded libraries were pooled at a concentration of 100 pM, and a 10 pM portion of the pool was added into an emulsion PCR-based template reaction mixture; in this reaction mixture, the fragments generated during library preparation were attached to Ion Sphere particles (ISPs) and clonally amplified. This process was carried out with the Ion One Touch 2 system and the Ion PGM Template OT2 200 kit. Quality control was performed on the Qubit 2.0 fluorometer with the Ion Sphere Quality Control assay. The optimal amount of library corresponds to the library dilution that gives template ISP percentages of 10 to 30%. The template-positive ISPs were then enriched and sequenced on the PGM with the Ion PGM Sequencing 200 kit v2 and an Ion 316 chip. Average coverage was determined by using an estimated genome size of 5.528 Mb and found to be 42× (range, 23 to 106×; see Table S1 in the supplemental material).

Bioinformatics. Sequence reads were mapped to the reference strain (Sakai, GenBank accession no. NC_002695) with BWA-MEM (24), and isolates with an average coverage of <20× were excluded from the analysis. Single nucleotide polymorphisms (SNPs) were then identified with GATK2 (25) in unified genotyper mode. Core genome positions (defined as sites for which a base was called for all isolates) that had a high-quality SNP (>90% consensus, minimum depth of 10×, genotype quality score of ≥30) or were the same as the reference base were extracted to produce a whole-core genome alignment for phylogenetic analysis. Sequence reads were also assembled *de novo* with the Torrent Suite Software, and the assemblies (contigs) were used for *in silico* analysis as described below.

***In silico* analysis.** Local BLAST databases were developed in either BioEdit (http://www.mbio.ncsu.edu/bioedit/bioedit.html) or BioNumerics v6.6 (Applied Maths) for multilocus sequence typing (MLST), *stx* subtyping, virulence gene detection, and lineage-specific polymorphism assay (LSPA-6) typing. The databases were then queried with the *de novo* assemblies. For MLST, the database consisted of the alleles of 15 housekeeping genes (*arcA*, *aroE*, *aspC*, *clpX*, *cyaA*, *dnaG*, *fadD*, *grpE*, *icdA*, *lysP*, *mdh*, *mtlD*, *mutS*, *rpoS*, *uidA*), which were downloaded from the EcMLST website (www.shigatox.net/ecmlst/cgi-bin/index); for *stx* subtyping, the database consisted of 600-bp sequences (342 bp of the C-terminal part of subunit A and 258 bp of the N-terminal part of subunit B) of 106 *stx* variants used by Scheutz et al. (23). For LSPA-6 typing, partial gene sequences of the six genes (*folD-sfmA*, *Z5935*, *yhcG*, *rbsB*, *rtcB*, and *arp-iclR*) were extracted from strain Sakai and used to form the database (26, 27). Isolates with genotype 111111 were classified as LSPA-6 lineage I, those with genotype 211111 were classified as LSPA-6 lineage I/II, and those with other derivations were classified as LSPA-6 lineage II.

Strains belonging to clade 8 (28) were identified on the basis of the discriminatory SNP (C/A at position 539 in gene ECs2357) described by Riordan et al. (29).

VirulenceFinder and ResFinder (http://cge.cbs.dtu.dk/services/) were used to determine the presence of *E. coli* virulence genes and acquired antibiotic resistance genes with identity thresholds of 85 and 98%, respectively (30, 31).

Data analysis. Ridom EpiCompare (http://www3.ridom.de/epicompare/) was used to calculate the discriminatory power of the typing methods. Odds ratios (ORs) were calculated with MedCalc (http://www.medcalc.org/calc/odds_ratio.php).

Recombination detection. The number of variable sites across the core genome alignment was calculated in sliding windows of 10,000 bp and plotted by using custom Python and R scripts to identify areas of the genome with an unusually high density of SNPs. The entire *E. coli* O157 core genome alignment was screened for recombination with BratNextGen (32) and also screened for recombination with BratNextGen when the two highly divergent and putative recombinant sequences (XH18570E and XH22083W) were excluded from the alignment.

Maximum-likelihood phylogenetic analysis. Maximum-likelihood core genome phylogenies were constructed with the RaxML software (33; Linux version 8) and PhyML (34). For the published RaxML phylogeny, the general time-reversible model of nucleotide substitution was used, with gamma-distributed rate heterogeneity across sites and 1,000 bootstrap replicates. Neighbor-joining phylogenies were also produced with MEGA version 5 (35). We confirmed that the tree topology was robust to the use of different software frameworks, as well as different choices of evolutionary model. In the absence of an obvious outgroup to O157, phylogenies were rooted at the midpoint between the two most divergent taxa in the trees.

Nucleotide sequence accession numbers. FASTQ sequences were deposited in the NCBI Sequence Read Archive under BioProject no. PRJNA283577 (http://www.ncbi.nlm.nih.gov/bioproject/PRJNA283577/).

RESULTS

Analysis of core genome. Our core genome alignment of the 105 *E. coli* O157 strains was 4,122,236 bp in length and contained 8,721 variable sites. However, a large number (4,110; 47%) of the SNPs were attributable to six "atypical" *E. coli* O157 strains, including a sorbitol-fermenting (SF) *E. coli* O157 strain (XH25052C) and three β-GUD-positive *E. coli* O157 strains (XH20443W, XH16200L, and XH24967A). When looking across the whole alignment, an ~150-kb region with a high density of SNPs (an average of 148.33 variable sites per 10,000-bp window, compared to a genome-wide average of 21.16 variable sites per 10,000-bp window) was observed near the end of the genome. Two of the atypical strains (XH18570E and XH22083W) were identified by the BratNextGen software as recombinant in this region, suggesting that genetic material has entered from a donor strain. See Fig. S1 in the supplemental material for plots of the distribution of variable sites across the genome in the presence and absence of recombinant strains XH18570E and XH22083W, as well as with and without recombinant regions excluded. Sanger sequencing of 792 bp of the *mutL* gene, which lies within the putative recombinant region, confirmed that the SNPs detected by WGS were genuine, and a BLAST analysis of this gene region showed that it was identical to the *mutL* gene of *E. coli* K-12.

Removal of recombinant strains XH18570E and XH22083W from the alignment, followed by the removal of additional, small, putative recombinant regions identified by BratNextGen for other strains, resulted in a 4,114,451-bp alignment containing 6,626 variable sites.

Epidemiological concordance. Eleven groups of two or more isolates known to be epidemiologically linked were present in our analysis, and the epidemiologically linked isolates formed clusters with 100% bootstrap support in the core genome phylogeny (Fig. 1). Within each epidemiologically linked cluster, there were fewer than four core SNPs separating the isolates. Three SNPs were identified among 10 isolates from a United Kingdom-wide phage type 8 (PT8) outbreak spanning 11 months in 2011. A maximum of two SNPs were identified among isolates from the other epidemiologically linked cases received by the SERL within a shorter time frame (0 to 14 days). The SNPs detected among epidemiologically related cases were confirmed by Sanger sequencing. For the SNP distances between all of the pairs of isolates, see Table S2 in the supplemental material. In three instances, isolates from cases with no apparent epidemiological link clustered together in the core SNP phylogeny and were separated by a maximum of 1 SNP. In two cases (XH11963F with XH12059K and XH15449Q with XH15521V), the isolates were temporally related (isolated within 8 days), had the same MLVA profile, and did not differ at any core SNP positions (i.e., their core genome sequences were identical), suggesting that there may have been an unidentified common source of infection. Another isolate (XH16734G) clustered with XH15449Q and XH15521V (differing from those isolates by one SNP) but was an MLVA double-locus variant (DLV) and was isolated ~4 months later. In the third case (XH11856D with XH12193B), the two isolates differed by one SNP and clustered in the core genome phylogeny with 100% bootstrap support, were MLVA single-locus variants (SLVs), and were isolated 39 days apart. Two additional MLVA SLVs (XH14013B with XH17884G and XH14653V with XH18908N) were observed in the data set; however, they were not related in time or space and differed by 36 and 126 SNPs, indicating that they were not closely related. The genetic variation among all pairs of isolates with no epidemiological links was 9 to 1,632 SNPs (see Table S2 in the supplemental material).

Concordance with MLVA and phage typing. A comparison of the minimum spanning tree produced by MLVA and the phylogeny based on the core genome showed that, although epidemiologically linked isolates clustered together in both trees, the overall topologies and the clustering of epidemiologically unrelated isolates differed between the trees. Within the core genome phylogeny, all of the sublineages we defined, and the majority of the clades within these were supported by bootstrap values of 100% (see Fig. S2 and S3 in the supplemental material). In contrast, the minimum spanning tree based upon the MLVA data (see Fig. S4 in the supplemental material) had relatively low bootstrap values, suggesting that MLVA provides meaningful phylogenetic relationships only for closely related isolates. Similarly, although phage typing was broadly congruent with the core genome phylogeny, identical PTs were distributed across the core genome phylogeny in different clades (Fig. 1). For example, PT4 was distributed in four of our defined sublineages across the tree, suggesting a lack of specificity of phage typing.

Discriminatory power of typing methods. The SNP data separated the 105 isolates into 81 different types (Simpson's index of diversity [SID], 0.989; 95% confidence interval [CI], 0.979 to 0.998), where isolates that differed by more than three SNPs were considered to be of different types. MLVA produced 80 different allelic profiles (SLVs were considered to be of the same type) and had the same discriminatory index as the SNP method (SID, 0.989; 95% CI, 0.979 to 0.998). The types were concordant, except for the two sets of MLVA SLV's that fell in different parts of the

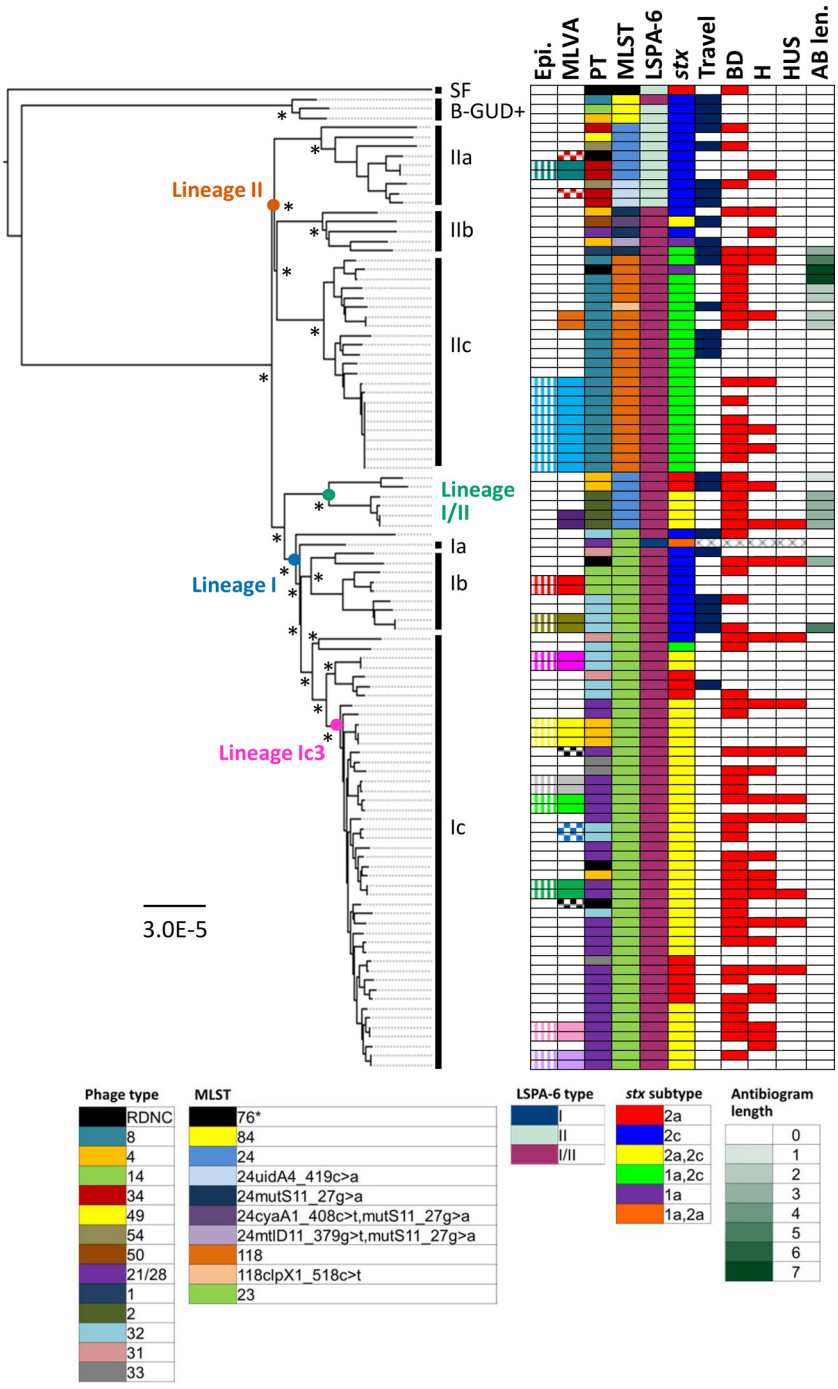

FIG 1 Maximum-likelihood core genome phylogeny for 103 *E. coli* O157 isolates from Lothian, Scotland. Recombinant sequences XH18570E and XH22083W were excluded. The tree was constructed with RAxML by using a general time-reversible model of nucleotide substitution and gamma distributed rate heterogeneity across sites. Branch lengths are in numbers of substitutions per site. Isolates known to be epidemiologically related (Epi) are in the same color. MLVA types shared by more than one isolate are indicated by filled boxes of the same color; SLV MLVA types are indicated by checkered boxes of the same color. PTs, multilocus STs, LSPA-6 types, *stx* subtypes, and the number of antibiotics (out of 15 tested) to which an isolate was resistant (AB len.) are shown. RDNC indicates that the PT reaction did not conform to a recognized pattern. Isolates associated with recent travel are in blue (Travel). All Lothian isolates were associated with diarrhea; isolates associated with bloody diarrhea (BD), hospitalization (H), and/or HUS are in red. Sublineages, defined as described in the text, are indicated by vertical black bars. One thousand bootstrap replicates were conducted, and bootstrap values of >98%, for nodes corresponding to the major lineages and sublineages described in the text, are indicated by asterisks.

core genome phylogeny and isolate XH16734G, which differed at two MLVA alleles (VNTR3 and VNTR9 by one repeat) from XH15449Q and XH15221V (Fig. 1; see Table S1 in the supplemental material). Phage typing produced 14 different lysis patterns and had a much lower SID than the other two methods (SID, 0.853; 95% CI, 0.817 to 0.89).

***stx* subtyping.** The *stx* subtypes were determined by PCR and *in silico* analyses. A relatively low diversity of subtypes was de-

tected among the 105 isolates (Fig. 1; see Table S1 in the supplemental material). The most common subtype was stx_{2a}/stx_{2c} ($n = 42$), followed by stx_{2c} only ($n = 26$) and stx_{1a}/stx_{2c} ($n = 24$). Eleven isolates carried stx_{2a} only, while two had stx_{1a} only.

The PCR results were concordant with the local BLAST and VirulenceFinder results for 102 (97%) and 97 (92%) isolates, respectively. The local BLAST search did not detect the stx_{2c} gene in three isolates found to carry stx_{2a} and stx_{2c} by PCR; in one case, only a partial sequence was identified, while in another, two different stx_{2a} subtypes were detected. VirulenceFinder produced similar results for these isolates; however, in one case, it detected the presence of an stx_{2c} subunit A but the adjacent subunit B was reported as stx_{2a}. Similarly, in four other cases, VirulenceFinder detected different stx_2 variants in subunits that lay beside each other in the genome. In one case, a stx_{2d} subunit A was reported in one isolate, which was not detected by PCR assay or the local BLAST search.

Virulence gene detection. VirulenceFinder and local BLAST searches were used to determine the presence of virulence genes; for a complete list of the genes detected in each isolate, see Table S1 in the supplemental material. A total of 23 virulence genes were detected among the 105 isolates. Twenty of the genes (*astA*, *eae*, *ehxA*, *espA*, *espB*, *espF*, *espJ*, *espP*, *etpD*, *gad*, *iha*, *iss*, *katP*, *nleA*, *nleB*, *nleC*, *prfB*, *tccP*, *tir*, and *toxB*) were present or partially present (<85% of the gene detected; see Table S1 in the supplemental material) in the majority (94%) of the isolates. In six isolates, one or more of these genes were not detected (see Table S1). These were the four "atypical" *E. coli* isolates (XH16200C, XH20443W, XH24967A, and XH25052C), one isolate that was negative for *espP* (XH12849W), and another that was negative for *iha* (XH14120C). The three remaining genes, *cdtB*, *cba*, and *celB*, were found in only a small proportion of the isolates.

The presence of *eae* and *hly* was detected by PCR in all of the study isolates; however, VirulenceFinder failed to detect these genes in 8 and 10 cases, respectively (see Table S1). The discordant results were further analyzed by performing local BLAST searches of these genes against the *de novo* assemblies. These revealed that the genes were present in all of the strains but were split among two or more different contigs, which explains why they were not detected by VirulenceFinder. Also, in some cases, part of the gene sequence was duplicated in the different contigs. For example, for XH11963F, bases 1 to 1602 of *eae* were present in contig c122 and bases 1388 to 2805 were present in contig c131, resulting in an overlap of 125 bp in the middle of the gene. The fractured and duplicated nature of these assemblies may be suggestive of low coverage in these areas, resulting in poor assembly.

Ten other genes not tested by PCR (*espF*, *espP*, *nleC*, *iha*, *katP*, *toxB*, *tccP*, *cba*, *celB*, *cdtB*) were further investigated by local BLAST searches, as they were not detected by VirulenceFinder in at least one of the isolates. The results of the local BLAST searches matched the VirulenceFinder results in the majority of the cases (313/413), where the genes were also not detected by the local BLAST searches. In 10 cases, only partial genes (<85% of the gene) were detected, while in 90 cases, the genes were present in multiple contigs, explaining why they were not detected by VirulenceFinder. Most notably, *tccP* was not detected in 75/105 cases because of its presence in multiple contigs (see Table S1 in the supplemental material). *tccP* can vary in length and contains proline-rich repeats (36), which are difficult to resolve by using short sequence reads, explaining its presence in multiple contigs and the failure of VirulenceFinder to detect this gene in most of the isolates.

Antimicrobial resistance gene detection. The isolates were tested for susceptibility to 15 different antimicrobial agents belonging to six different antimicrobial classes. A relatively low level of resistance was detected (Fig. 1; see Table S1 in the supplemental material); 15 (14%) out of 105 isolates showed resistance to one or more antibiotics. Streptomycin ($n = 14$), sulfamethoxazole ($n = 14$), and tetracycline ($n = 9$) were the antibiotics to which resistance was most frequently observed. Two isolates were resistant to six antibiotics: ampicillin, kanamycin, sulfamethoxazole, streptomycin, tetracycline, and trimethoprim. We did not observe an association between antimicrobial resistance and foreign travel (OR, 1.2; $P = 0.7753$). There were no disagreements between the observed and predicted susceptibility patterns; however, phenotypic and genotypic resistance to several antibiotics (ampicillin, sulfamethoxazole, streptomycin, trimethoprim-sulfamethoxazole, and trimethoprim) was observed in one isolate that was epidemiologically linked to a strain in which no resistance was detected (XH18778Q and XH18795X; see Table S1 in the supplemental material). This may be related to the loss or gain of an unstable plasmid carrying the resistance genes.

MLST. Ten different multilocus sequence types (STs) were identified among the 105 strains (Fig. 1). ST23 was the most common type ($n = 56$), followed by ST118 ($n = 22$) and ST24 ($n = 14$). The three β-GUD-positive strains belonged to ST84 (ST65 seven-locus scheme), while the SF *E. coli* O157 strain was ST76, which corresponds to previous reports (37). Five new STs were identified, as they differed by at least one SNP, in one or more alleles, from those present in the EcMLST database. Also, in four strains, a single nucleotide gap/no call was detected in one of the alleles (see Table S1 in the supplemental material). Inspection of the sequence reads at these positions showed that the base was present in only about one-third of the reads and so would have been excluded during minimum variant frequency filtering. As expected, the ST correlated well with the core genome phylogeny.

Evolutionary analysis of *E. coli* O157. Figure 1 shows the maximum-likelihood phylogeny constructed from the core genome alignment of the O157 strains. The atypical strains (SF and β-GUD positive) cluster together and are phylogenetically distinct from the β-GUD-negative *E. coli* O157 strains isolated from the majority of Lothian patients in recent years. The rest of the phylogeny can be divided into three clades that correspond to recently described lineages I, I/II, and II (38). As noted by Dallman et al., LSPA-6 typing does not adequately resolve these lineages; the *folD-sfmA* polymorphism does not differentiate lineages I and II. The lineages can be further divided into sublineages with distinct genotypic and phenotypic characteristics. As indicated in Fig. 1, the bootstrap values for all of the lineages and sublineages were greater than 98% (with the majority being 100%), indicating a high level of support for the clusters observed. A strong concordance was noted between the lineages defined in the phylogeny and the clusters inferred from the proportion of shared ancestry tree in the BratNextGen analysis (see Fig. S2 in the supplemental material). Sublineage IIa consists of ST24 strains, or ST24 SLVs, that carry stx_{2c} only and consist of PT34, PT49, and PT54. Sublineage IIb is a heterogeneous group with respect to the PT and *stx* profile, while sublineage IIc consists of ST118 PT8 strains that have acquired stx_{1a}. Fifty-seven percent of the strains associated with antimicrobial resistance (antibiogram length of >1) were

present within lineage IIc. Lineage I/II consists of ST24 PT4 and PT2 strains carrying either stx_{2a} only or both stx_{2a} and stx_{2c}, respectively, and belonging to clade 8 (see Table S1 in the supplemental material) that have been associated with large outbreaks of severe disease in the United States (28). Lineage I strains belong to ST23 and consist of three main sublineages (Ia, Sakai; Ib, stx_{2c}-only genotype; and Ic) that can be further subdivided into three main clades. The largest, named Ic3 here, consists predominately of PT21/28 stx_{2a} and stx_{2c} strains that are responsible for the majority of the human infections reported in Scotland and the United Kingdom (5). Of note, no isolates in this clade were associated with acquired antimicrobial resistance or foreign travel.

Correlation of genotype with disease severity. The enhanced surveillance data showed that all 105 (100%) patients had diarrhea, 64 (61%) had bloody diarrhea, 30 (29%) were hospitalized, and 10 (9%) developed HUS. Six (60%) of the 10 HUS cases were in children <16 years old (see Table S1 in the supplemental material). No deaths were recorded.

Seven (70%) of the 10 HUS cases were associated with genotype Ic3. Statistical analysis of the results showed a significant association between Ic3 and HUS (OR, 4.5938; $P = 0.0351$). Ic3 was also significantly associated with hospitalization (OR, 3.1503; $P = 0.0103$) but not bloody diarrhea (OR, 2.1212; $P = 0.0826$). No significant correlations between the other genotypes and clinical outcomes were observed; however, some tests were based upon very small numbers of samples, making it difficult to determine statistically meaningful results.

Ic3 strains carry stx_{2a} (and usually stx_{2c}), and previous reports have suggested that stx_{2a} is associated with increased virulence and more severe human disease (39, 40). We found that 8 (80%) of 10 isolates from HUS cases carried the stx_{2a} gene; however, statistical analysis showed that the presence of stx_{2a} was not significantly associated with HUS (OR, 4.4444; $P = 0.0678$) or hospitalization (OR, 2.0829; $P = 0.0986$), suggesting that other features of 1c3 and/or host factors are likely to be important in the development of severe complications of infection.

DISCUSSION

Two of the most important features of a typing method are its ability to distinguish between unrelated strains and to correctly classify epidemiologically related isolates from an outbreak or cluster as part of the same clone (41). This study showed that the identification of core genome SNPs following WGS provided a highly discriminatory method for subtyping of *E. coli* O157, giving results concordant with the epidemiological data. The method was also concordant with MLVA for the identification of epidemiologically linked cases, even though different areas and a much smaller percentage of the genome were targeted by MLVA, demonstrating the value of MLVA for outbreak detection. A recent study carried out by Public Health England (PHE) showed that MLVA was as sensitive as WGS for identifying linked cases of *E. coli* O157 infection but, because of the time taken to determine the relatedness of MLVA DLVs occasionally observed in larger outbreaks, found that WGS resolved all of the cases in a cluster faster (42). In this study, among the isolates with no known epidemiological links, there were three discordant results between MLVA and WGS. In one case, an MLVA DLV was shown to differ by only one SNP from two other isolates (that were identical by WGS and MLVA), while in two cases, the WGS data showed that MLVA SLVs were not, in fact, closely related in terms of the core genome. We currently use MLVA profiles (exact matches and SLVs) to alert health protection teams and HPS of putative linked cases, which are then further investigated for epidemiological links. Together, the data suggest that WGS may provide more precise data and reduce the unnecessary deployment of health protection resources in the investigation of putatively linked isolates. Turabelidze et al. (14) showed that a cluster of *E. coli* O157 cases associated with the consumption of romaine lettuce and salad bar exposures was better resolved by using core genome SNPs than by using MLVA and PFGE.

The determination of SNP distances between strains showed that some diversity existed among isolates from epidemiologically linked cases. We found three or fewer SNP differences among isolates from known epidemiologically linked cases. Definition of variation is important for the assessment of relatedness and establishment of a targeted approach to case inclusion in outbreaks and contact tracing. Others have reported SNP diversity among epidemiologically related cases. Dallman et al. (42) found five or fewer SNP differences among 183 epidemiologically related isolates from different clusters received by PHE, Underwood et al. (17) detected four SNP differences among 16 isolates of *E. coli* O157 isolated during a farm outbreak, and Joensen et al. detected seven SNP differences among six outbreak cases (30). Various factors may explain the degree of genomic variation reported in different outbreaks, including the strains sampled, outbreak duration, the sequencing platforms used, and the criteria used to define high-quality SNPs.

In the past, considerable effort has been made to standardize *E. coli* typing methods to enable comparisons of data produced in different laboratories (e.g., PulseNet, CDC). The standardization of MLVA between reference laboratories in the United Kingdom has been invaluable for the rapid detection of cross-border outbreaks (22), and this study has confirmed its utility in outbreak detection. It is likely that the standardization of data produced by WGS in different laboratories will be a major challenge and require considerable collaboration between clinical and reference laboratories.

A major advantage of WGS is its ability to produce a well-supported, high-resolution phylogeny and therefore an appropriate method for understanding or tracking the evolutionary relationships between strains and for detailed epidemiological surveillance of *E. coli* O157. Although the isolates sequenced in this study were from a single Scottish health board, phylogenetic analysis revealed a population structure remarkably consistent with that in previous studies, in particular, work recently carried out at PHE, where a large number of strains of *E. coli* O157 ($n = 1,075$) from clinical and animal sources collected over a 29-year period in the United Kingdom were analyzed (38). All lineages and sublineages, except Ia, were represented in our strain collection. Similar to the PHE study, the main clusters within lineages I, I/II, and II consisted of isolates belonging to the PTs commonly associated with human disease in the United Kingdom, PT21/28, PT2, and PT8, respectively.

Another major advantage of WGS is the additional information that can be extracted, including important virulence and antibiotic resistance determinants. We were able to correctly infer antibiotic susceptibility profiles from the WGS data, suggesting that WGS is a suitable alternative to current routine laboratory testing for the surveillance of antimicrobial resistance and the detection of emerging resistance phenotypes. Resistance to strepto-

mycin, sulfamethoxazole, and tetracycline was most often observed, consistent with previous reports (43, 44, 45), most likely because of the selection pressure imposed by the use of these agents in clinical and veterinary medicine.

Using the genes present in the VirulenceFinder database (76 genes plus variants), we found limited variation in gene content among the isolates. Differences were observed mostly among the atypical isolates because of the carriage of different plasmids. For example, ST76 SF *E. coli* O157 did not carry *espP*, *katP*, and *toxB*, which is consistent with the carriage of the pSFO157 plasmid (46). This strain also did not carry *espF*, *iha*, and *tccP* but harbored *cdtB*. Eklund et al. (47) found most SF *E. coli* strains to possess the gene cluster (*cdtV*) encoding the cytolethal distending toxin family, which includes *cdtB*. Among the NSF *E. coli* O157 isolates, we found *cdtB* exclusively within sublineage Ib, so it may be suitable for use as a marker for this sublineage.

Although the virulence genes were detected with a high degree of accuracy, some discordant results were observed, mostly related to the limitation of short sequence read assembly to resolve repeat regions and, in some cases, low read depth resulting in poor assemblies. The sensitivity of *in silico* detection software, such as VirulenceFinder, depends upon the use of high-quality assemblies, and assembly efficiency is dependent on read depth across the whole genome. Jünemann et al. (48) compared benchtop sequencers to show that the optimal coverage (based on N50 values) with sequencing data generated by the PGM and MiSeq was ~40- and 75-fold, respectively, while Desai et al. (49) reported that the coverage required to produce a good assembly was dependent on the assembly software used; however, 50× was optimal for most of the widely used software (e.g., Velvet, SOAPdenovo, and AbySS) to assemble an *E. coli* genome with Illumina reads. Other gene-finding software (e.g., Ridom SeqSphere) suggests a minimum coverage of ≥50× for PGM data and ≥75× for MiSeq data (Ridom GmbH, Münster, Germany). In this study, the average read depth ranged from 23 to 106 (see Table S1 in the supplemental material); therefore, in some cases, optimal coverage was not obtained, explaining, in part, the failure of VirulenceFinder to detect some genes. However, a gene-finding method that does not consider each contig separately may make better use of the available sequencing information.

The VirulenceFinder database also contained the *stx* genes as individual subunit sequences rather than holotoxins; however, the combination of both subunit sequences is essential for accurate differentiation of the subtypes, explaining why, in some cases, two subunits lying adjacent to each other were reported as different *stx* subtypes. It is essential that the *stx* genes can be accurately detected and characterized, as they play an essential role in the development of disease and some studies (38, 39, 40) have found that certain subtypes (e.g., stx_{2a}) have been associated with increased pathogenicity. WGS has the advantage of being able to detect multiple *stx* genes of the same subtype, which is not possible by PCR. For example, in one strain, we detected two different stx_{2a} gene variants. Recently, Ashton et al. (50) used a combined mapping and *de novo* approach to detect multiple genes of the same *stx* subtype in ~15% of the isolates tested. In this study, we may have underestimated the number of isolates with multiple genes of the same subtype, as the bioinformatics approach used may not have detected these. The significance of the presence of multiple genes of the same *stx* subtype is unknown and needs further investigation.

It is currently not possible to determine MLVA types from WGS data because of short read lengths; however, with advances in technology such as single-molecule real-time sequencing (e.g., PacBio) (51) and nanopore sequencing (52), which provide longer read lengths capable of spanning tandem repeat regions, this information will be captured and will be available to help define the evolutionary relationships of isolates. The molecular basis of phage typing is also not known, so it is currently not possible to extract PT from WGS data. MLST types were successfully determined *in silico*. Although the 15-locus scheme provides a low level of resolution, there are various softwares (Applied Maths, Ridom SeqSphere, BIGSdb) available for whole-genome MLST, enabling the analysis of thousands of genes. It is thought that whole-genome MLST will be easier to standardize than core SNP analysis and enable a hierarchical approach to be used (53); however, the method requires allele curation and may be less discriminatory.

The development of improved software to overcome the limitations of short read assembly is required to realize the potential of WGS for in *silico* data analysis. Precise knowledge and understanding of the genome contents of the strains causing clinical infections should enable the identification of risk factors for improved patient management. We have shown that Ic3 strains were significantly associated with severe disease; however, contrary to recent work (38), we did not find carriage of stx_{2a} to be essential for the development of HUS. Further in-depth analysis of these Ic3 strains is required to determine features likely to be important in the development of severe disease.

Overall, we have demonstrated that WGS offers the potential to streamline reference laboratory processes by the use of a single diagnostic tool to generate the information required to support the clinical management of cases and public health investigations and interventions to control disease spread. It has the potential to transform the way we assess the relatedness of strains and the risk of development of severe complications. However, issues relating to ease of performance and standardization, as well as information technology infrastructure and data storage, need to be addressed before it is introduced routinely.

ACKNOWLEDGMENTS

M.W. was supported by the Evolution and Transfer of Antibiotic Resistance (EvoTAR) project (European Commission Framework Programme 7) and by a research fellowship as part of a Wellcome Trust Strategic grant to the Centre for Immunity, Infection and Evolution (grant reference no. 095831). We thank the Edinburgh and Lothian Health Foundation for funding this work.

We thank HPS (National Services Scotland) as our commissioners; Lindsey Davis, health protection nurse, NHS Lothian; Lynda Browning, epidemiologist, HPS; and Jon Manning and Al Ivens for bioinformatics services.

REFERENCES

1. **Karmali MA, Petric M, Lim C, Fleming PC, Arbus GS, Lior H.** 1985. The association between hemolytic uremic syndrome and infection by verotoxin-producing *Escherichia coli*. J Infect Dis **151:**775–782. http://dx.doi.org/10.1093/infdis/151.5.775.
2. **Tarr PI, Gordon CA, Chandler WL.** 2005. Shiga-toxin-producing *Escherichia coli* and haemolytic uraemic syndrome. Lancet **365:**1073–1086. http://dx.doi.org/10.1016/S0140-6736(05)71144-2.
3. **Pollock K.** 18 August 2005. Enhanced surveillance of haemolytic uraemic syndrome and other thrombotic microangiopathies in Scotland, 2003–2004. Euro Surveill **10:**E050519.5. http://www.eurosurveillance.org/ViewArticle.aspx?ArticleId=2708.
4. **Browning L, Allison L, Hanson M, Hawkins G.** 2014. VTEC in Scotland:

enhanced surveillance, reference laboratory and clinical reporting data. Health Protection Scotland, Glasgow, Scotland. http://www.hps.scot.nhs.uk/documents/ewr/pdf2015/1533.pdf
5. **Locking ME, Allison LA, Rae L, Hanson M.** 14 May 2014. VTEC and HUS in Scotland, 2013: enhanced surveillance, reference laboratory and clinical reporting data. Health Protection Scotland, Glasgow, Scotland. http://www.hps.scot.nhs.uk/ewr/article.aspx.
6. **Sodha SV, Heiman K, Gould H, Bishop R, Iwamoto M, Swerdlow DL, Griffin PM.** 2015. National patterns of *Escherichia coli* O157 infections, USA, 1996–2011. Epidemiol Infect **143:**267–273. http://dx.doi.org/10.1017/S0950268814000880.
7. **Roberts JA, Upton PA, Azene G.** 2000. *Escherichia coli* O157:H7; an economic assessment of an outbreak. J Public Health Med **22:**99–107. http://dx.doi.org/10.1093/pubmed/22.1.99.
8. **Cowden JM, Ahmed S, Donaghy M, Riley A.** 2001. Epidemiological investigation of the central Scotland outbreak of *Escherichia coli* O157 infection, November to December 1996. Epidemiol Infect **126:**335–341.
9. **Pennington TH.** 2014. E. coli O157 outbreaks in the United Kingdom: past, present, and future. Infect Drug Resist **7:**211–222.
10. **Locking ME, O'Brien SJ, Reilly WJ, Wright EM, Campbell DM, Coia JE, Browning LM, Ramsay CN.** 2001. Risk factors for sporadic cases of *Escherichia coli* O157 infection: the importance of contact with animal excreta. Epidemiol Infect **127:**215–220.
11. **Howie H, Mukerjee A, Cowden J, Leith J, Reid T.** 2003. Investigation of an outbreak of *Escherichia coli* O157 infection caused by environmental exposure at a scout camp. Epidemiol Infect **131:**1063–1069. http://dx.doi.org/10.1017/S0950268803001250.
12. **Walker TM, Ip CL, Harrell RH, Evans JT, Kapatai G, Dedicoat MJ, Eyre DW, Wilson DJ, Hawkey PM, Crook DW, Parkhill J, Harris D, Walker AS, Bowden R, Monk P, Smith EG, Peto TE.** 2013. Whole-genome sequencing to delineate *Mycobacterium tuberculosis* outbreaks: a retrospective observational study. Lancet Infect Dis **13:**137–146. http://dx.doi.org/10.1016/S1473-3099(12)70277-3.
13. **Roetzer A, Diel R, Kohl TA, Rückert C, Nübel U, Blom J, Wirth T, Jaenicke S, Schuback S, Rüsch-Gerdes S, Supply P, Kalinowski J, Niemann S.** 2013. Whole genome sequencing versus traditional genotyping for investigation of a *Mycobacterium tuberculosis* outbreak: a longitudinal molecular epidemiological study. PLoS Med **10:**e1001387. http://dx.doi.org/10.1371/journal.pmed.1001387.
14. **Turabelidze G, Lawrence SJ, Gao H, Sodergren E, Weinstock GM, Abubucker S, Wylie T, Mitreva M, Shaikh N, Gautom R, Tarr PI.** 2013. Precise dissection of an *Escherichia coli* O157:H7 outbreak by single nucleotide polymorphism analysis. J Clin Microbiol **51:**3950–3954. http://dx.doi.org/10.1128/JCM.01930-13.
15. **Price J, Gordon NC, Crook D, Llewelyn M, Paul J.** 2013. The usefulness of whole genome sequencing in the management of *Staphylococcus aureus* infections. Clin Microbiol Infect **19:**784–789. http://dx.doi.org/10.1111/1469-0691.12109.
16. **den Bakker HC, Allard MW, Bopp D, Brown EW, Fontana J, Iqbal Z, Kinney A, Limberger R, Musser KA, Shudt M, Strain E, Wiedmann M, Wolfgang WJ.** 2014. Rapid whole-genome sequencing for surveillance of *Salmonella enterica* serovar enteritidis. Emerg Infect Dis **20:**1306–1314. http://dx.doi.org/10.3201/eid2008.131399.
17. **Underwood AP, Dallman T, Thomson NR, Williams M, Harker K, Perry N, Adak B, Willshaw G, Cheasty T, Green J, Dougan G, Parkhill J, Wain J.** 2013. Public health value of next-generation DNA sequencing of enterohemorrhagic *Escherichia coli* isolates from an outbreak. J Clin Microbiol **51:**232–237. http://dx.doi.org/10.1128/JCM.01696-12.
18. **Dallman TJ, Byrne L, Launders N, Glen K, Grant KA, Jenkins C.** 2015. The utility and public health implications of PCR and whole genome sequencing for the detection and investigation of an outbreak of Shiga toxin-producing *Escherichia coli* serogroup O26:H11. Epidemiol Infect **143:**1672–1680. http://dx.doi.org/10.1017/S0950268814002696.
19. **Outhred AC, Jelfs P, Suliman B, Hill-Cawthorne GA, Crawford AB, Marais BJ, Sintchenko V.** 2015. Added value of whole-genome sequencing for management of highly drug-resistant TB. J Antimicrob Chemother **70:**1198–1202.
20. **Public Health England.** 30 September 2011. UK *E. coli* O157 outbreak associated with soil on vegetables. Public Health England, London, United Kingdom. http://www.hpa.org.uk/NewsCentre/NationalPressReleases/2011PressReleases/110930Ecolioutbreakassocwithsoilonveg/.
21. **Ahmed R, Bopp C, Borczyk A, Kasatiya S.** 1987. Phage-typing scheme for *Escherichia coli* O157:H7. Infect Dis **155:**806–809. http://dx.doi.org/10.1093/infdis/155.4.806.
22. **Holmes A, Perry N, Willshaw G, Hanson M, Allison L.** 2015. Inter-laboratory comparison of multi-locus variable-number tandem repeat analysis (MLVA) for verocytotoxin-producing *Escherichia coli* O157 to facilitate data sharing. Epidemiol Infect **143:**104–107. http://dx.doi.org/10.1017/S0950268814000739.
23. **Scheutz F, Teel LD, Beutin L, Pierard D, Buvens G, Karch H, Mellmann A, Caprioli A, Tozzoli R, Morabito S, Strockbine 425 NA, Melton-Celsa AR, Sanchez M, Persson S, O'Brien AD.** 2012. Multicenter evaluation of a sequence-based protocol for subtyping Shiga toxins and standardizing Stx nomenclature. J Clin Microbiol **50:**2951–2963. http://dx.doi.org/10.1128/JCM.00860-12.
24. **Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup.** 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics **25:**2078–2079.
25. **DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM, Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ.** 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet **43:**491–498. http://dx.doi.org/10.1038/ng.806.
26. **Yang Z, Kovar J, Kim J, Nietfeldt J, Smith DR, Moxley RA, Olson ME, Fey PD, Benson AK.** 2004. Identification of common subpopulations of non-sorbitol-fermenting, beta-glucuronidase-negative *Escherichia coli* O157:H7 from bovine production environments and human clinical samples. Appl Environ Microbiol **70:**6846–6854. http://dx.doi.org/10.1128/AEM.70.11.6846-6854.2004.
27. **Ziebell K, Steele M, Zhang Y, Benson A, Taboada EN, Laing C, McEwen S, Ciebin B, Johnson R, Gannon V.** 2008. Genotypic characterization and prevalence of virulence factors among Canadian *Escherichia coli* O157:H7 strains. Appl Environ Microbiol **74:**4314–4323. http://dx.doi.org/10.1128/AEM.02821-07.
28. **Manning SD, Motiwala AS, Springman AC, Qi W, Lacher DW, Ouellette LM, Mladonicky JM, Somsel P, Rudrik JT, Dietrich SE, Zhang W, Swaminathan B, Alland D, Whittam TS.** 2008. Variation in virulence among clades of *Escherichia coli* O157:H7 associated with disease outbreaks. Proc Natl Acad Sci U S A **105:**4868–4873. http://dx.doi.org/10.1073/pnas.0710834105.
29. **Riordan JT, Viswanath SB, Manning SD, Whittam TS.** 2008. Genetic differentiation of *Escherichia coli* O157:H7 clades associated with human disease by real-time PCR. J Clin Microbiol **46:**2070–2073. http://dx.doi.org/10.1128/JCM.00203-08.
30. **Joensen KG, Scheutz F, Lund O, Hasman H, Kaas RS, Nielsen EM, Aarestrup FM.** 2014. Real-time whole-genome sequencing for routine typing, surveillance, and outbreak detection of verotoxigenic *Escherichia coli*. J Clin Microbiol **52:**1501–1510. http://dx.doi.org/10.1128/JCM.03617-13.
31. **Zankari E, Hasman H, Cosentino S, Vestergaard M, Rasmussen S, Lund O, Aarestrup FM, Larsen MV.** 2012. Identification of acquired antimicrobial resistance genes. J Antimicrob Chemother **67:**2640–2644. http://dx.doi.org/10.1093/jac/dks261.
32. **Marttinen P, Hanage WP, Croucher NJ, Connor TR, Harris SR, Bentley SD, Corander J.** 2012. Detection of recombination events in bacterial genomes from large population samples. Nucleic Acids Res **40:**e6. http://dx.doi.org/10.1093/nar/gkr928.
33. **Stamatakis A.** 2014. RAxML Version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics **30:**1312–1313. http://dx.doi.org/10.1093/bioinformatics/btu033.
34. **Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O.** 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol **59:**307–321. http://dx.doi.org/10.1093/sysbio/syq010.
35. **Tamura K, Peterson D, Peterson N, Stecher G, Nei M, andKumar S.** 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol **28:**2731–2737. http://dx.doi.org/10.1093/molbev/msr121.
36. **Garmendia J, Ren Z, Tennant S, Midolli Viera MA, Chong Y, Whale A, Azzopardi K, Dahan S, Sircili MP, Franzolin MR, Trabulsi LR, Phillips A, Gomes TA, Xu J, Robins-Browne R, Frankel G.** 2005. Distribution of *tccP* in clinical enterohemorrhagic and enteropathogenic *Escherichia coli*

isolates. J Clin Microbiol **43:**5715–5720. http://dx.doi.org/10.1128/JCM.43.11.5715-5720.2005.

37. **Rump LV, Meng J, Strain EA, Cao G, Allard MW, Gonzalez-Escalona N.** 2012. Complete DNA sequence analysis of enterohemorrhagic *Escherichia coli* plasmid pO157_2 in β-glucuronidase-positive *E. coli* O157:H7 reveals a novel evolutionary path. J Bacteriol **194:**3457–3463. http://dx.doi.org/10.1128/JB.00197-12.
38. **Dallman T, Ashton P, Byrne L, Perry N, Petrovska L, Ellis R, Allison L, Hanson M, Holmes A, Gunn G, Chase-Topping M, Woolhouse M, Grant K, Gally D, Wain J, Jenkins.** 27 July 2015. Applying phylogenomics to understand the emergence of Shiga toxin producing *Escherichia coli* O157:H7 strains causing severe human disease in the United Kingdom. Microb Genomics http://dx.doi.org/10.1099/mgen.0.000029.
39. **Persson S, Olsen KE, Ethelberg S, Scheutz F.** 2007. Subtyping method for *Escherichia coli* Shiga toxin (verocytotoxin) 2 variants and correlations to clinical manifestations. J Clin Microbiol **45:**2020–2024. http://dx.doi.org/10.1128/JCM.02591-06.
40. **Bielaszewska M, Mellmann A, Bletz S, Zhang W, Köck R, Kossow A, Prager R, Fruth A, Orth-Höller D, Marejková M, Morabito S, Caprioli A, Piérard D, Smith G, Jenkins C, Curová K, Karch H.** 2013. Enterohemorrhagic *Escherichia coli* O26:H11/H-: a new virulent clone emerges in Europe. Clin Infect Dis **56:**1373–1381. http://dx.doi.org/10.1093/cid/cit055.
41. **Struelens M.** 1998. Molecular epidemiologic typing systems of bacterial pathogens: current issues and perspectives. Mem Inst Oswaldo Cruz **93:**581–585. http://dx.doi.org/10.1590/S0074-02761998000500004.
42. **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.** 2015. Whole genome sequencing for national surveillance of Shiga toxin producing *Escherichia coli* O157. Clin Infect Dis **61:**305–312. http://dx.doi.org/10.1093/cid/civ318.
43. **Meng J, Zhao S, Doyle MP, Joseph SW.** 1998. Antibiotic resistance of *Escherichia coli* O157:H7 and O157:NM isolated from animals, food, and humans. J Food Prot **61:**1511–1514.
44. **Schroeder CM, Zhao C, DebRoy C, Torcolini J, Zhao S, White DG, Wagner DD, McDermott PF, Walker RD, Meng J.** 2002. Antimicrobial resistance of *Escherichia coli* O157 isolated from humans, cattle, swine, and food. Appl Environ Microbiol **68:**576–581. http://dx.doi.org/10.1128/AEM.68.2.576-581.2002.
45. **Vidovic S, Tsoi S, Medihala P, Liu J, Wylie JL, Levett PN, Korber DR.** 2013. Molecular and antimicrobial susceptibility analyses distinguish clinical from bovine *Escherichia coli* O157 strains. J Clin Microbiol **51:**2082–2088. http://dx.doi.org/10.1128/JCM.00307-13.
46. **Brunder W, Karch H, Schmidt H.** 2006. Complete sequence of the large virulence plasmid pSFO157 of the sorbitol-fermenting enterohemorrhagic *Escherichia coli* O157:H- strain 3072/96. Int J Med Microbiol **296:**467–474. http://dx.doi.org/10.1016/j.ijmm.2006.05.005.
47. **Eklund M, Bielaszewska M, Nakari UM, Karch H, Siitonen A.** 2006. Molecular and phenotypic profiling of sorbitol-fermenting *Escherichia coli* O157:H$^-$ human isolates from Finland. Clin Microbiol Infect **12:**634–641. http://dx.doi.org/10.1111/j.1469-0691.2006.01478.x.
48. **Jünemann S, Sedlazeck FJ, Prior K, Albersmeier A, John U, Kalinowski J, Mellmann A, Goesmann A, von Haeseler A, Stoye J, Harmsen D.** 2013. Updating benchtop sequencing performance comparison. Nat Biotechnol **31:**294–296. http://dx.doi.org/10.1038/nbt.2522.
49. **Desai A, Marwah VS, Yadav A, Jha V, Dhaygude K, Bangar U, Kulkarni V, Jere A.** 2013. Identification of optimum sequencing depth especially for de novo genome assembly of small genomes using next generation sequencing data. PLoS One **8:**e60204. http://dx.doi.org/10.1371/journal.pone.0060204.
50. **Ashton P, Perry N, Ellis RJ, Petrovska L, Wain J, Grant K, Jenkins C, Dallman T.** 2015. Insight into Shiga toxin genes encoded by *Escherichia coli* O157 from whole genome sequencing. PeerJ **3:**e739. http://dx.doi.org/10.7717/peerj.739.
51. **Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B, Bibillo A, Bjornson K, Chaudhuri B, Christians F, Cicero R, Clark S, Dalal R, Dewinter A, Dixon J, Foquet M, Gaertner A, Hardenbol P, Heiner C, Hester K, Holden D, Kearns G, Kong X, Kuse R, Lacroix Y, Lin S, Lundquist P, Ma C, Marks P, Maxham M, Murphy D, Park I, Pham T, Phillips M, Roy J, Sebra R, Shen G, Sorenson J, Tomaney A, Travers K, Trulson M, Vieceli J, Wegener J, Wu D, Yang A, Zaccarin D, Zhao P, Zhong F, Korlach J, Turner S.** 2009. Real-time DNA sequencing from single polymerase molecules. Science **323:**133–138. http://dx.doi.org/10.1126/science.1162986.
52. **Schneider GF, Dekker C.** 2012. DNA sequencing with nanopores. Nat Biotechnol **30:**326–328. http://dx.doi.org/10.1038/nbt.2181.
53. **Maiden MC, Jansen van Rensburg MJ, Bray JE, Earle SG, Ford SA, Jolley KA, McCarthy ND.** 2013. MLST revisited: the gene-by-gene approach to bacterial genomics. Nat Rev Microbiol **11:**728–736. http://dx.doi.org/10.1038/nrmicro3093.