AMERICAN SOCIETY FOR MICROBIOLOGY | Journal of Clinical Microbiology®

EPIDEMIOLOGY

Validation of Whole-Genome Sequencing for Identification and Characterization of Shiga Toxin-Producing *Escherichia coli* To Produce Standardized Data To Enable Data Sharing

Anne Holmes,[a] Timothy J. Dallman,[b] Sharif Shabaan,[c] Mary Hanson,[a] Lesley Allison[a]

[a]Scottish *E. coli* O157/STEC Reference Laboratory (SERL), Royal Infirmary of Edinburgh, Edinburgh, Scotland

[b]Gastrointestinal Bacteria Reference Unit (GBRU), Reference Microbiology Services, National Infection Service, Public Health England, London, United Kingdom

[c]Centre for Infectious Diseases, University of Edinburgh, The Roslin Institute, Easter Bush, United Kingdom

ABSTRACT Whole-genome sequencing (WGS) is rapidly becoming the method of choice for outbreak investigations and public health surveillance of microbial pathogens. The combination of improved cluster resolution and prediction of resistance and virulence phenotypes provided by a single tool is extremely advantageous. However, the data produced are complex, and standard bioinformatics pipelines are required to translate the output into easily interpreted epidemiologically relevant information for public health action. The main aim of this study was to validate the implementation of WGS at the Scottish *Escherichia coli* O157/STEC Reference Laboratory (SERL) using the Public Health England (PHE) bioinformatics pipeline to produce standardized data to enable interlaboratory comparison of results generated at two national reference laboratories. In addition, we evaluated the BioNumerics whole-genome multilocus sequence typing (wgMLST) and *E. coli* genotyping plug-in tools using the same data set. A panel of 150 well-characterized isolates of Shiga toxin-producing *E. coli* (STEC) that had been sequenced and analyzed at PHE using the PHE pipeline and database (SnapperDB) was assembled to provide identification and typing data, including serotype (O:H type), sequence type (ST), virulence genes (*eae* and Shiga toxin [*stx*] subtype), and a single-nucleotide polymorphism (SNP) address. To validate the implementation of sequencing at the SERL, DNA was reextracted from the isolates and sequenced and analyzed using the PHE pipeline, which had been installed at the SERL; the output was then compared with the PHE data. The results showed a very high correlation between the data, ranging from 93% to 100%, suggesting that the standardization of WGS between our reference laboratories is possible. We also found excellent correlation between the results obtained using the PHE pipeline and BioNumerics, except for the detection of stx_{2a} and stx_{2c} when these subtypes are both carried by strains.

KEYWORDS Shiga toxin-producing *Escherichia coli*, whole-genome sequencing

Shiga toxin-producing *Escherichia coli* (STEC) is recognized as an important pathogen due to the serious and sometimes fatal complications that may occur following infection. Key virulence factors of STEC include intimin (encoded by the *eaeA* gene) enabling the pathogen to attach to epithelial cells, and the Shiga toxins (Stx1 and Stx2), which inhibit protein synthesis and can cause cell death (1). Hemolytic-uremic syndrome (HUS) is the most severe complication of STEC infection and most commonly occurs in children under 5 years of age (2–4). Scotland has one of the highest incidence rates of HUS in the world (3.4×10^5 in children <5 years), and although *E. coli* serotype O157:H7 is the most common cause, other serotypes are also implicated (5).

Transmission occurs via the fecal-oral route and is most often associated with the

Received 11 September 2017 **Returned for modification** 8 October 2017 **Accepted** 12 December 2017

Accepted manuscript posted online 20 December 2017

Citation Holmes A, Dallman TJ, Shabaan S, Hanson M, Allison L. 2018. Validation of whole-genome sequencing for identification and characterization of Shiga toxin-producing *Escherichia coli* to produce standardized data to enable data sharing. J Clin Microbiol 56:e01388-17. https://doi.org/10.1128/JCM.01388-17.

Editor Alexander Mellmann, University Hospital Münster

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

consumption of contaminated food or water, direct or indirect contact with animals (especially ruminants) and their feces, or person-to-person spread (6). Although most cases are sporadic, outbreaks do occur (7–9), and control strategies are essential to limit the spread of this pathogen. In Scotland, enhanced surveillance of STEC, introduced in 1999 for *E. coli* O157 and extended to non-O157 and cases with HUS in 2003, facilitates the rapid identification of outbreaks and monitoring of the epidemiology and clinical outcome of this important pathogen.

Various typing methods have been employed over the years to detect outbreaks, including phage typing, pulsed-field gel electrophoresis (PFGE), and multilocus variable-number tandem-repeat analysis (MLVA). More recently, studies have shown that whole-genome sequencing (WGS) offers benefits over these techniques in terms of improved resolution and the additional useful information (e.g., virulence and resistance genes) that can be easily obtained from the sequencing data (10–14). As a result, clinical and reference laboratories worldwide are implementing this technology for routine use. The Gastrointestinal Bacterial Reference Unit (GBRU) at Colindale, Public Health England (PHE), which provides reference services for STEC isolated in England and Wales, replaced MLVA of *E. coli* O157 and phenotypic serotyping of non-O157 *E. coli* isolates with WGS in October 2015 (15). More recently, the Centers for Disease Control and Prevention (CDC) reported the implementation of WGS for the characterization and typing of STEC in the United States (16).

One of the main challenges of WGS is data standardization, which is converting the raw information into meaningful results that can be easily communicated and compared. PHE has developed a semiautomated bioinformatics pipeline using free open-source software, which provides hierarchical typing information, including a single-nucleotide polymorphism (SNP) address (numerical code describing the population structure at seven different SNP thresholds), providing a nomenclature that can be easily shared (15). The CDC has validated the use of whole-genome multilocus sequence typing (wgMLST) and the *E. coli* genotyping plug-in tool using the commercial platform BioNumerics (Applied Maths), which has the advantage of combining quality assessment, data analysis, and databasing of metadata in a single user-friendly platform (16).

It is important that the reference laboratories in the United Kingdom can easily compare data for national surveillance and detect cross-border outbreaks. The main aim of this study was to validate the implementation of WGS at the SERL using the PHE bioinformatics pipeline to produce standardized data to enable interlaboratory comparisons of results between our reference laboratories. In addition, we evaluated the wgMLST and *E. coli* genotyping plug-in tools in BioNumerics to facilitate comparison with others using this approach. Recently, PulseNet International reported wgMLST as the method of choice for the surveillance of foodborne bacterial pathogens (17).

MATERIALS AND METHODS

Isolates. The study panel comprised 110 *E. coli* O157 and 40 non-O157 *E. coli* isolates obtained between 16 September 2014 and 2 February 2017. The strains were representative of STEC isolates detected in human samples in Scotland and the United Kingdom but included four non-O157 *E. coli* isolates from milk ($n = 1$), cheese ($n = 1$), and venison sausages ($n = 2$). The isolates had been previously characterized using traditional methods (including phage typing, PCR, and MLVA) at the SERL and represented the diversity of Scottish *E. coli* strains causing clinical infection (see Table S1 in the supplemental material). Furthermore, the isolates had been previously sequenced and analyzed by PHE using a validated laboratory procedure and bioinformatics pipeline (15, 18) to provide identification and typing data, including serotype (O:H type), virulence gene detection (*eaeA*, *bfpA*, *aggR*, *ipaH*, *aaiC*, *itcA*, *sta1*, and *stb*), 7-gene sequence typing (MLST), and an SNP address. To validate the implementation of WGS at the SERL, the isolates were reextracted, resequenced, and analyzed at the SERL, and the results were compared with those generated by PHE to identify and address any issues hindering the comparison of data between our laboratories. The data set was also used to evaluate the BioNumerics wgMLST and *E. coli* genotyping plug-in tools (Applied Maths).

Whole-genome sequencing. DNA was extracted on the QIAsymphony using the DSP DNA minikit (Qiagen, Crawley, UK) following a prelysis step, as recommended by the manufacturer. The purity and quantity of the DNA were measured with a NanoDrop ND-1000 (NanoDrop Products, Thermo Scientific) and a Qubit fluorimeter 3.0 (Thermo Fisher Scientific) with the double-stranded DNA (dsDNA) assay HS kit.

Libraries were prepared using the Nextera XT DNA kit, according to the manufacturer's instructions (Illumina, Cambridge, UK). Paired-end sequencing was performed on the Illumina MiSeq using either the v2 or v3 reagent kit. Isolates were sequenced in 11 runs, and the cluster density ranged from 919 K/mm^2 to 1,981 K/mm^2 (mean, 1,435 K/mm^2), the passing filter (PF) ranged from 77.05% to 95.44% (mean, 88.30%), and the average Q30 ranged from 78.05 to 94.10% (mean, 88.72%) (Table S1, see "Run metrics"). In the first four runs, the index primer combination N704/S507 was associated with low coverage; therefore, these primers were not used in subsequent runs. Samples with an average coverage of $<40\times$ underwent repeat sequencing.

Data analysis. (i) PHE pipeline. Fastq files were processed using the PHE bioinformatics pipeline and database SnapperDB (https://github.com/phe-bioinformatics/snapperdb) for *E. coli* O157, as previously described, with some modifications (15). Briefly, reads were quality trimmed, with bases with a Phred score below 30 removed from the trailing edge using Trimmomatic (19). Kraken (20) was used to identify bacterial species and determine if samples were mixed. Using the GeneFinder tool (M. Doumith, unpublished data), FASTQ reads were mapped to a panel of serotype (21) and virulence genes using Bowtie 2 (22), and the best match to each target was reported with metrics, including coverage, depth, mixture, and nucleotide similarity in XML format for quality assessment. Only *in silico* predictions of serotype and virulence that matched to a gene determinant at >80% nucleotide identity over >80% target gene length were accepted. MLST alleles (7 gene) were determined using MOST (23). Shiga toxin gene subtyping was performed using a combined mapping and BLAST approach, as previously described (24). Isolates of *E. coli* O157 were then reference mapped to Sakai GenBank accession no. BA000007.2 using BWA and GATK2 to identify variants, as previously described (25). Pairwise SNP comparison and single-linkage clustering at 7 different levels of increasing similarity (SNP thresholds, 250, 100, 50, 25, 10, 5, and 0) were performed to produce an "SNP address." This clustering results in a discrete seven-digit code, where each number represents the cluster membership at each descending SNP distance threshold. The SnapperDB installed at the SERL contained approximately 3,000 SNP addresses produced previously at PHE, which included the 109 *E. coli* O157:H7 isolates in the validation panel. Reads were also *de novo* assembled using SPAdes 3.6.1 (26). The observed quality parameters for each sample are shown in Table S1. The average coverage was determined for those isolates that underwent reference-based assembly (*E. coli* O157:H7 only) and ranged from 41× to 158× (mean, 84×), and the N_{50} >1,000 scores for all samples ranged from 60,462 bp to 179,895 bp (mean, 128,422 bp).

(ii) BioNumerics version 7.6. Fastq files from sequenced genomes were processed using the BioNumerics calculation engine and the wgMLST client plug-in. Assembly-free and assembly-based allele detection analyses were performed for each isolate. The assembly was performed using SPAdes integrated into the calculation engine, and basic assembly metrics were calculated (Table S1). The average read coverage for all samples ranged from 41× to 218× (mean, 103×), and the N_{50} >300 scores ranged from 57,549 bp to 179,899 bp (mean, 119,330 bp). The assembled genomes were then analyzed using the *E. coli* genotyping plug-in, which contains databases for serotype, virulence, and resistance prediction obtained from the Center for Genomic Epidemiology (DTU, Lyngby, Denmark [https://cge.cbs.dtu.dk/services/data.php]). The detection parameters for gene detection using BLAST were set to 90% sequence identity and 60% sequence coverage. The *E. coli* genotyping plug-in also has an *in silico* PCR tool for the detection of virulence genes and Shiga toxin gene subtypes using previously published primers (27–29). The *in silico* PCR settings were set to allow for 1 mismatch in the primer sequence binding sites. cgMLST (2,513 core loci synchronized with EnteroBase schema) dendrograms were produced using categorical differences with no scaling factor and complete linkage cluster analysis.

(iii) Traditional laboratory methods. Latex agglutination for the O157 antigen was performed using the Oxoid Wellcolex kit, and API20E was used for biochemical confirmation of *E. coli*. Phage typing was performed with 16 phages, as previously described (30). Antibiotic sensitivity patterns were determined using the disk 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. ATCC 25922 was used as a control strain. Intermediate (I) (Table S1) phenotypes were counted as susceptible. Discordant results between the WGS and *in vitro* susceptibility results were repeated. Using the phenotypic results as a gold standard, sensitivity was calculated by dividing the number of isolates that were genotypically resistant by the total number of isolates exhibiting resistance phenotypes. Specificity was calculated by dividing the number of isolates that were genotypically susceptible by the total number of isolates with susceptible phenotypes. Nalidixic acid was not included in the analysis, as resistance to this antibiotic occurs via chromosomal mutation(s), which were not investigated.

MLVA was performed as previously described (31). Raw data (.fsa files) from an ABI 3130 genetic analyzer (Applied Biosystems) were imported and analyzed in BioNumerics version 7.6 (Applied Maths, Sint-Martens-Latem, Belgium) with the MLVA plug-in. Real-time PCR was used for the detection of stx_1, stx_2, *rfbO157*, and *eaeA* (32). Real-time PCRs (RT-PCRs) contained 1× QuantiTect PCR mix (Qiagen UK Ltd., Crawley, UK), 0.2 μM each primer set, 0.1 μM each probe, and 2 μl DNA template in a final volume of 20 μl. The in-house primers and probe for the detection of stx_{2f} were included in the multiplex PCR: stx2f-F, TTGTCACAGTGATAGCAGAAGCTCTG; stx2f-R, CAGTTCAGGGTAAGGTCAACATCC; and stx2f-P, FAM-CGCTGTCTGAGGCATCTCCGCTTTATAC. Amplification was carried out on the CFX (Bio-Rad Laboratories Ltd.) with an initial denaturation at 95°C for 15 min, followed by 45 cycles of 95°C for 15 s and 60°C for 1 min, with data collection. Data were analyzed using the CFX Manager software.

TABLE 1 Concordance between SERL and PHE results

Characteristic tested	No. of isolates			Concordance (%) between PHE and SERL results
	Routine[a]	PHE	SERL	
Species identification	150	150	150	100
O:H serotype	110	150	150	100
Sequence type		150	150	100
eaeA detected	128	127	127	100
Other virulence genes detected[b]		1	1	100
stx subtype		150	149	99
SNP address		109[c]	101	93

[a]Traditional testing using API20E for *E. coli* species identification, latex agglutination for O157 antigen detection only, and real-time PCR for the detection of *eaeA*.
[b]Detection of *bfpA*, *aggR*, *ipaH*, *aaiC*, *itcA*, *sta1*, and/or *stb*.
[c]SNP addresses were determined only for the 109 isolates of *E. coli* O157:H7.

Accession number(s). FASTQ sequences were deposited in the NCBI Sequence Read Archive under BioProject no. PRJNA419720.

RESULTS

Validation of the PHE bioinformatics pipeline implemented at the SERL. (i) Serotype determination. The validation panel included 22 different O serogroups, four "O group unidentifiable," and 16 H serogroups, and there was complete agreement between the results obtained by the SERL and reported by PHE (Tables 1 and S1). In two instances (isolates from epidemiologically related cases), different O128 variants were detected for *wxy* and *wzy*, O128ab and O128ac, respectively. However, this has been reported previously (21), and it was suggested that a distinction between these variants is no longer necessary due to their high genetic similarity and lack of evidence to suggest a significant difference in clinical outcome associated with these variants. As a result, the isolates were named based on the best *wzx* match, O128ab. Also, in two isolates of O113:H4, O113:H17 was initially reported by the SERL; however, this was due to an error in the reference database, where an H4 gene was labeled as H17.

(ii) Seven-gene MLST. The MLST profiles were 100% concordant. Six novel STs and five novel alleles were detected in the data set (Table S1).

(iii) Virulence gene detection. For the detection of the *eaeA* gene, there was 100% agreement between the SERL and PHE WGS results; however, in one case, the pipeline failed to detect *eaeA* in an isolate that was positive by PCR. This was because the coverage and identity were 74.6% and 72.66%, respectively, which were lower than the threshold settings of the Gene Finder tool. All isolates were negative for the other seven virulence genes, except for one isolate that was positive for *sta1* (no. 146) (Table S1). There was 99% agreement with *stx* subtyping. In one case, an *E. coli* O157 isolate (no. 46) (Table S1), which was positive for stx_{2a} and stx_{2c} by PHE and confirmed by real-time PCR by the SERL (data not shown), was reported as having stx_{2a} only. This was due to low-quality data in the gene region, and upon repeat testing (subculture, DNA extraction, and sequencing), the isolate was found to be positive for both stx_{2a} and stx_{2c}.

(iv) SNP address. SNP addresses were determined for the isolates of *E. coli* O157:H7, and identical 7-digit profiles were obtained for 101/109 (93%) isolates, showing the WGS and data analysis to be highly reproducible. In the eight discrepant cases, only the last digit of the SNP address differed, which means they fall in the same 5-SNP cluster, and further analysis showed they differed by either 1 ($n = 5$), 2 ($n = 1$), or 3 ($n = 2$) SNPs. The coverages for the discrepant cases were 41, 66, 72, 83, 87, 98, 118, and 131×, suggesting the differences were not related to low coverage. Also, no differences were found when the PHE FASTQs were analyzed with the SERL pipeline, suggesting the discrepancies were related to the SERL resequencing procedure.

BioNumerics analysis and correlation with the PHE pipeline. (i) Serotype determination. There was 100% correlation between O:H types (Table 2). Similar to the PHE pipeline results, different O128 variants for *wzx* and *wzy* were reported in isolates from two epidemiologically related cases (no. 136 and 137) (Table S1) using BioNumerics.

TABLE 2 Concordance between PHE pipeline and BioNumerics (*in silico* PCR and BLAST) results for gene detection

Characteristic tested	No. of isolates				Concordance (%)[a]
	PCR	PHE pipeline	BioNumerics *in silico* PCR	BLAST[b]	
Species identification		150	150		100
O:H serotype		150		150	100
H7 *fliC*		110	110		100
eaeA	128	127[c]	127[d]	128	99.2
MLST		150		147	98
stx_1	54	54	54	54	100
stx_{1a}		47	47	46	100
stx_{1c}		7	7	7	100
stx_2	139	139	134	139	100
stx_{2a}		71	35	69	97.2
stx_{2b}		15	15	14	100
stx_{2c}		87	45	46	52.9
stx_{2d}		2	2	2	100
stx_{2f}		8	8	8	100
stx_{2g}		1	1	1	100

[a]Concordance between the PHE pipeline and combined *in silico* PCR and BLAST BioNumerics results.
[b]All genes/subunits were taken into account, e.g., stx_1, $stxA_1$, $stxB_1$, $stxA_2$, and $stxB_2$ (see Table S1 in the supplemental material).
[c]*eaeA* was not detected in an *E. coli* O145:H34 isolate using the PHE pipeline but was positive for *eaeA* by traditional PCR and *in silico* PCR.
[d]*In silico* PCR did not detect *eaeA* in an *E. coli* O34:H4 isolate due to a 3-bp mismatch in the reverse primer binding sequence.

(ii) Virulence gene detection. *(a)* In silico *PCR*. The results of virulence gene detection are shown in and S1. All isolates were detected by *in silico* PCR for the *E. coli* species-specific genes and H7 *fliC*. A total of 127 of 128 (99%) isolates were positive for *eaeA* by *in silico* PCR. In one O34:H4 isolate (Table S1, no. 114), *eaeA* was not detected by *in silico* PCR due to three mismatches (in bold) in the reverse primer binding sequence (CCACCTGC**T**GCAAC**G**A**C**AGG). However, *eaeA* was detected in an O145:H34 isolate (Table S1, no. 111) positive for *eaeA* by PCR but not detected with the PHE pipeline.

In total, 54 isolates were positive for stx_1; all of these were detected by *in silico* PCR, and there was 100% agreement with stx_1 subtyping. Forty-seven isolates carried stx_{1a} alone or in combination with stx_{2c} and/or stx_{2a}, while the other seven isolates carried stx_{1c} alone or in combination with stx_{2b}. No isolates were positive for stx_{1d}.

In total, 134 of 139 (96.4%) isolates were positive by *in silico* PCR for stx_2. Further analysis showed stx_2 was not detected in five isolates (Table S1, no. 23, 36, 37, 76, and 98), because either the primer binding regions fell into different contigs (Table S1, no. 23, 36, and 37) or the reverse primer was not detected (Table S1, no. 76 and 98).

There was good correlation with stx_2 subtyping, except for the detection of stx_{2a} and stx_{2c} in isolates carrying both of these variants. Fifteen isolates were positive for stx_{2b}, alone or in combination with stx_{1a} or stx_{1c}, two isolates were positive for stx_{2d}, eight isolates were positive for stx_{2f}, and one isolate was positive for stx_{2g}. Only 35 of 71 (49.3%) isolates were positive for stx_{2a}, alone or in combination with stx_{2c} and/or stx_{1a}, and 45 of 87 (51.7%) isolates were stx_{2c} positive alone or in combination with stx_{2a} or stx_{1a}. In all discrepant cases, stx_{2a} and/or stx_{2c} were not detected in strains carrying both of these variants. Changing the settings to increase the number of primer mismatches to 2, 3, or 4 did not significantly alter the results.

(b) BLAST. There was 100% correlation with the detection of *eaeA*, stx_1, and stx_2 when all genes were taken into consideration (Table S1). There was good agreement between stx_1 gene subtyping; in one case, stx_1 was not differentiated as an stx_{1a} variant (Table S1, no. 118). There was also good agreement for stx_2 subtyping, except for the detection of stx_{2a} and stx_{2c} in strains carrying both of these variants; only 46 of 87 (53%) stx_{2c} genes were detected in these strains. Changing the settings to reduce the identity

and coverage (e.g., to 50% and 10%, respectively) did not significantly alter the results. In two cases, no stx_{2a} gene was identified in strains carrying stx_{2a} and stx_{2c}; in one case, an stx_2 gene was not differentiated as variant b, and in two cases, stx_{2g} was detected in isolates not known to carry this gene.

(iii) Seven-gene MLST. Seven-digit MLST profiles were obtained for 147/150 (98%) isolates; one ($n = 2$) or two ($n = 1$) loci were missing in three strains (Table S1).

(iv) Comparison of MLVA, SNP analysis, and cgMLST. Core-genome MLST (cgMLST) dendrograms were created for the *E. coli* O157 and non-O157 *E. coli* isolates, shown in Fig. S1 and S2, respectively. For the *E. coli* O157 isolates, 18 groups (1 to 18, Fig. S1) with two or more isolates that were temporally linked and considered highly related by MLVA (exact match or single-locus variant) were present in the data set. Of these groups, 13 had isolates with the same SNP address (0- or 5-SNP threshold), suggesting that the cases were linked. In four of the five other groups (2, 10, 14, and 18, Fig. S1), the isolates had different SNP addresses, indicating that the cases were not closely related. In one group (16, Fig. S1), four of the five isolates had the same SNP address (0- or 5-SNP threshold), while one isolate differed by 4 digits (50-SNP threshold), suggesting it was not related to the other four cases.

The cgMLST analysis showed that the 13 groups of isolates considered to be putatively related by MLVA and SNP analysis also clustered using cgMLST, with a maximum of three loci separating the isolates. Of the remaining five groups, four produced data consistent with the SNP analysis, suggesting the isolates were not linked; however, in one case (2), only a single cgMLST locus separated the isolates. Further analysis showed these isolates differed by only 8 SNPs. Overall, the number of loci (cgMLST) that differed between unrelated isolates was lower than the number of SNPs (e.g., 22 loci versus 43 SNPs [10], 11 loci versus 22 SNPs [14], 9 loci versus 21 SNPs [18], and 41 loci versus 76 SNPs [16], respectively), indicating that cgMLST is less discriminatory than SNP analysis.

For the non-O157 *E. coli* isolates (Fig. S2), there were four groups with two isolates each known to be epidemiologically related, and the data showed that these epidemiologically linked isolates clustered by cgMLST. Within each cluster, there was a maximum of 1 allele separating the isolates. The sporadic isolates clustered based on sequence type and serogroup and differed by $\geq$9 loci.

Prediction of antimicrobial resistance phenotype using the ResFinder database in BioNumerics. Of the 150 isolates tested in this study, 16 (14 *E. coli* O157 and 2 non-O157 *E. coli* [11%] isolates) showed phenotypic resistance to one or more antibiotics, and 15 isolates were found to carry resistance gene(s) (Table 3 and S1). A total of 5 mismatches were observed between the phenotypic and genotypic data, resulting in a concordance of 99.8% (2,095/2,100 resistance tests), sensitivity of 95.3% (41/43 tested positive), and specificity of 99.9% (2,054/2,057 tested negative). In one isolate (no. 124, an *E. coli* O165:H25), phenotypic resistance to ampicillin and amoxicillin-clavulanate was observed, but no resistance genes were detected. In two isolates, *catA1* was detected, but phenotypic resistance to chloramphenicol was not observed; and in one isolate, *aph(3′)-Ia* was detected, but resistance to kanamycin was not detected. Phenotypic resistance to nalidixic acid was observed in two isolates, but the resistance mechanisms (e.g., *gyrA* mutations) were not investigated. Two isolates carried the *mph*(B) gene, suggesting macrolide resistance.

DISCUSSION

There is no doubt that WGS offers significant advantages over traditional methods for the characterization and typing of bacterial pathogens. However, there are major challenges associated with implementing WGS in clinical and reference laboratories. Foremost, the data produced are complex and require specialist bioinformatics skills and knowledge for processing and analysis. Many laboratories, however, do not have this expertise and require automated user-friendly bioinformatics workflows that provide reproducible standardized data that are easily shared between laboratories.

PHE is at the forefront of using WGS for disease surveillance and outbreak detection

TABLE 3 Comparison of phenotypic antimicrobial resistance with resistance gene detection[a]

Antimicrobial	No. of isolates with phenotypic resistance	No. of isolates with resistance gene(s) detected	Resistance gene combinations (no. of isolates)
Ampicillin	4	3	bla_{TEM-1B} (3)
Amoxicillin-clavulanate	4	3	bla_{TEM-1B} (3)
Chloramphenicol	1	3	*catA1*, *cmlA1*, *cmI* (1)
			catA1 (2)
Kanamycin	1	2	*aph(3′)-Ia* (2)
Streptomycin	15	15	*straA*, *staB* (8)
			strA, *strB*, *aadA1* (4)
			strA, *strB*, *aadA1*, *aadA2* (1)
			aadA12 (2)
Tetracycline	13	13	*tet*(A) (9)
			tet(B) (4)
Trimethoprim-sulfamethoxazole	5	5	*sul1*, *sul2*, *dfrA1* (4)
			sul1, *sul3*, *dfrA1* (1)
Nalidixic acid	3	NT	
Sulfamethoxazole	NT	10	*sul1* (2)
			sul2 (8)
Macrolide	NT	2	*mph*(B) (2)

[a]$n = 16$. NT, not tested.

and has developed validated protocols and bioinformatics pipelines for a number of bacterial pathogens, including STEC (15, 18, 33, 34). The SERL works closely with the GBRU, along with their respective public health teams and national surveillance centers, to rapidly identify, monitor, and prevent UK-wide outbreaks. It is therefore important that the data produced in our laboratories can be easily compared and communicated. The results of this study showed that the data produced by the SERL and those reported by PHE were highly concordant, ranging from 93% to 100%. Most importantly, following repeat subculture and DNA extraction, sequencing, and analysis, the majority (93%) of the SNP addresses were identical. In a small number of cases, the last digit of the SNP address was different, which was related to differences of between 1 and 3 SNPs. Mutations arising from repeat subculture and/or variations in the laboratory protocol and sequencing platform may explain discrepancies, as it was not possible to completely standardize the laboratory protocol; e.g., PHE used the Illumina HiSeq, whereas the MiSeq was used by the SERL. Importantly, however, these small differences fell within the 5-SNP cutoff used to categorize isolates as being putatively related. Previous work showed that STEC cases with known epidemiological links had either no difference in their core genomes or fell within a 5-SNP threshold (15, 35).

The installation of the pipeline at the SERL required bioinformatics expertise; however, once installed, the pipeline is semiautomated and can be run using a limited number of commands. It is important that the software versions and reference databases employed are kept up to date and synchronized between laboratories. Moreover, to ensure that strains are given the same SNP address, it is essential that SnapperDB is populated with the same set of strains, and, ideally, a single instance of SnapperDB would be used to provide a uniform nomenclature. Our laboratories are currently working closely to achieve this.

In the SERL, we have considerable experience using BioNumerics, having previously used this software for PFGE and MLVA (31). The data set was therefore used to evaluate the recently developed plug-in tools for wgMLST and *E. coli* genotyping. The advantages of this system are that no previous bioinformatics experience is required, and high-performance computers are not necessary, as data processing is carried out on an external calculation engine. There is, however, a cost associated with using the calculation engine of ~9 euros per sample for wgMLST. The results showed very good correlation between the data produced using the PHE pipeline and those with Bio-

Numerics, despite the different analysis methods employed. The exception was for *stx* subtype prediction, where the *in silico* PCR/BLAST approach in BioNumerics was not ideally suited for the sensitive detection of the highly related *stx* subtypes stx_{2a} and stx_{2c} when both of these variants were carried by a strain. We also analyzed some of these strains using VirulenceFinder 1.5 (https://cge.cbs.dtu.dk/services/VirulenceFinder/) using the assembled genome/contigs option and threshold settings of 90% and 60%, respectively, for identity and minimum length; we found the same results as those obtained using BioNumerics (data not shown). This poses a problem for our laboratory, as the most common STEC lineage in the United Kingdom, lineage 1c (PT21/28), most often carries both of these variants (25, 35). As a result, we plan to use the PHE pipeline for the prediction of *stx* subtype, which is based on a combined mapping and BLAST approach and has been shown to correlate very well (99%) with conventional PCR for the detection of *stx* variants, including those containing multiple subtypes (24). Recently, Lindsey et al. (16) evaluated BioNumerics for the identification and characterization of STEC in the United States and showed good concordance between *stx* subtyping and PCR; however, only a relatively small number of strains ($n = 19$) were evaluated, and none carried both the stx_{2a} and stx_{2c} genes.

BioNumerics analysis using the cgMLST approach discriminated the isolates with a resolution comparable to that of the core-genome SNP analysis. The number of loci differing between unrelated strains was lower than the number of SNPs detected, showing the method to be slightly less discriminatory. However, this is expected, considering that a smaller proportion of the genome is being sampled. Similar findings have been reported with other pathogens (36). A major advantage of cgMLST is that the same set of alleles can be used for all *E. coli* clones, whereas for SNP analysis, different databases and reference genomes are required for different clones. Furthermore, the method is easily scalable, and the allele numbering system is amenable to standardization and the development of a nomenclature system. It is for these reasons, among others, that PulseNet International has recently reported wgMLST as their current method of choice for the analysis of WGS data (17).

The surveillance of antimicrobial resistance (AMR) to monitor trends and emerging resistances of STEC is important. Similar to previous work conducted in our laboratory, we found that relatively low numbers of isolates carried antimicrobial resistance genes (35). The most common resistance genes detected were those for streptomycin, tetracycline, and sulfamethoxazole. Studies by others have also found a similar relatively low level of resistance among STEC isolates. Day et al. (37) found that 82.6% of STEC O157 and 64.7% of STEC O26 isolates were predicted to be fully susceptible to 11 diverse classes of antimicrobials.

We found good agreement between phenotypic and WGS predicted antimicrobial susceptibility, similar to that reported by Lindsey et al. (16); this provides further evidence to support the replacement of phenotypic testing for WGS-inferred susceptibility for surveillance purposes, as the results do not directly influence individual patient management. However, further work is required to incorporate the detection of resistance due to mutational events into our bioinformatics workflow, and it will be essential to update resistance databases as new mechanisms are discovered.

The findings of this study show that the standardization of SNP-based WGS and analysis between our reference laboratories is possible. The method was highly reproducible, providing confidence in the accuracy of the results. Others have reported the high accuracy and reproducibility of WGS-based bacterial typing (38, 39). Kozyreva et al. (38) used a panel of different bacterial species and developed a WGS workflow to provide highly accurate, sensitive, specific, and reproducible results. Furthermore, they have established a framework of "best practices" for the quality management of both "wet-lab" and "dry-lab" components to help others validate WGS. Mellmann et al. (39) conducted a ring trial involving five laboratories using a standardized methodology and observed highly concordant results. This work demonstrated the feasibility of an external quality assessment system for WGS, which is an important quality assurance activity to help ensure the validity of test results and for accreditation. The Global

Microbial Identifier has organized a proficiency testing program for three different pathogens, including *E. coli*, in which the SERL is participating. In addition, the laboratory will participate in the cluster analysis component of the forthcoming EQA-8 scheme for the typing of STEC organized by the European Food- and Waterborne Diseases and Zoonoses Programme of the European Centre for Disease Control and Prevention (ECDC). The SERL implemented WGS-based typing for STEC in August 2017 following successful validation, and it is currently undergoing United Kingdom Accreditation Service (UKAS) accreditation.

SUPPLEMENTAL MATERIAL

Supplemental material for this article may be found at https://doi.org/10.1128/JCM.01388-17.

SUPPLEMENTAL FILE 1, PDF file, 0.1 MB.
SUPPLEMENTAL FILE 2, PDF file, 0.2 MB.
SUPPLEMENTAL FILE 3, XLS file, 0.5 MB.

ACKNOWLEDGMENT

We thank NHS National Services Scotland for supporting this work.

REFERENCES

1. Law D. 2000. Virulence factors of *Escherichia coli* O157 and other Shiga toxin-producing *E. coli*. J Appl Microbiol 88:729–745. https://doi.org/10.1046/j.1365-2672.2000.01031.x.
2. Tarr PI, Gordon CA, Chandler WL. 2005. Shiga-toxin-producing *Escherichia coli* and haemolytic uraemic syndrome. Lancet 365:1073–1086. https://doi.org/10.1016/S0140-6736(05)71144-2.
3. Pollock K. 2005. Enhanced surveillance of haemolytic uraemic syndrome and other thrombotic microangiopathies in Scotland, 2003–2004. Euro Surveill 10:E050519.5. https://doi.org/10.2807/esw.10.20.02708-en.
4. Gould LH, Demma L, Jones TF, Hurd S, Vugia DJ, Smith K, Shiferaw B, Segler S, Palmer A, Zansky S, Griffin PM. 2009. Hemolytic uremic syndrome and death in persons with *Escherichia coli* O157:H7 infection, foodborne diseases active surveillance network sites, 2000–2006. Clin Infect Dis 49:1480–1485. https://doi.org/10.1086/644621.
5. Lynn RM, O'Brien SJ, Taylor CM, Adak GK, Chart H, Cheasty T, Coia JE, Gillespie IA, Locking ME, Reilly WJ, Smith HR, Waters A, Willshaw GA. 2005. Childhood hemolytic uremic syndrome, United Kingdom and Ireland. Emerg Infect Dis 14:590–596. https://doi.org/10.3201/eid1104.040833.
6. 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. https://doi.org/10.1017/S0950268801006045.
7. 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. https://doi.org/10.1017/S0950268801005520.
8. Sodha SV, Lynch M, Wannemuehler K, Leeper M, Malavet M, Schaffzin J, Chen T, Langer A, Glenshaw Hoefer MD, Dumas N, Lind L, Iwamoto M, Ayers T, Nguyen T, Biggerstaff M, Olson C, Sheith A, Braden C. 2011. Multistate outbreak of *Escherichia coli* O157:H7 infections associated with a national fast-food chain, 2006: a study incorporating epidemiological and food source traceback results. Epidemiol Infect 139:309–316. https://doi.org/10.1017/S0950268810000920.
9. Jenkins C, Dallman TJ, Launders N, Willis C, Byrne L, Jorgensen F, Eppinger M, Adak GK, Aird H, Elviss N, Grant KA, Morgan D, McLauchlin J. 2015. Public health investigation of two outbreaks of Shiga toxin-producing *Escherichia coli* O157 associated with consumption of watercress. Appl Environ Microbiol 81:3946–3952. https://doi.org/10.1128/AEM.04188-14.
10. Eppinger M, Mammel MK, Leclerc JE, Ravel J, Cebula TA. 2011. Genomic anatomy of *Escherichia coli* O157:H7 outbreaks. Proc Natl Acad Sci U S A 108:20142–20147. https://doi.org/10.1073/pnas.1107176108.
11. 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. https://doi.org/10.1128/JCM.01696-12.
12. 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. https://doi.org/10.1128/JCM.01930-13.
13. 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. https://doi.org/10.1128/JCM.03617-13.
14. Rusconi B, Sanjar F, Koenig SS, Mammel MK, Tarr PI, Eppinger M. 2016. Whole genome sequencing for genomics-guided investigations of *Escherichia coli* O157:H7 outbreaks. Front Microbiol 7:985. https://doi.org/10.3389/fmicb.2016.00985.
15. 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. https://doi.org/10.1093/cid/civ318.
16. Lindsey RL, Pouseele H, Chen JC, Strockbine NA, Carleton HA. 2016. Implementation of whole genome sequencing (WGS) for identification and characterization of Shiga toxin-producing *Escherichia coli* (STEC) in the United States. Front Microbiol 7:766. https://doi.org/10.3389/fmicb.2016.00766.
17. Nadon C, Van Walle I, Gerner-Smidt P, Campos J, Chinen I, Concepcion-Acevedo J, Gilpin B, Smith AM, Man Kam K, Perez E, Trees E, Kubota K, Takkinen J, Nielsen EM, Carleton H, FWD-NEXT Expert Panel. 2017. PulseNet International: vision for the implementation of whole genome sequencing (WGS) for global food-borne disease surveillance. Euro Surveill 22:pii=30544. https://doi.org/10.2807/1560-7917.ES.2017.22.23.30544.
18. Chattaway MA, Dallman TJ, Gentle A, Wright MJ, Long SE, Ashton PM, Perry NT, Jenkins C. 2016. Whole genome sequencing for public health surveillance of Shiga toxin-producing *Escherichia coli* other than serogroup O157. Front Microbiol 7:258. https://doi.org/10.3389/fmicb.2016.00258.
19. Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120. https://doi.org/10.1093/bioinformatics/btu170.
20. Wood DE, Salzberg SL. 2014. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol 15:R46. https://doi.org/10.1186/gb-2014-15-3-r46.
21. Joensen KG, Tetzschner AM, Iguchi A, Aarestrup FM, Scheutz F. 2015. Rapid and easy *in silico* serotyping of *Escherichia coli* isolates by use of

whole-genome sequencing data. J Clin Microbiol 53:2410–2426. https://doi.org/10.1128/JCM.00008-15.
22. Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10:R25. https://doi.org/10.1186/gb-2009-10-3-r25.
23. Tewolde R, Dallman T, Schaefer U, Sheppard CL, Ashton P, Pichon B, Ellington M, Swift C, Green J, Underwood A. 2016. MOST: a modified MLST typing tool based on short read sequencing. PeerJ 4:e2308. https://doi.org/10.7717/peerj.2308.
24. 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. https://doi.org/10.7717/peerj.739.
25. 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 C. 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 Genom 1:e000029. https://doi.org/10.1099/mgen.0.000029.
26. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol 19:455–477. https://doi.org/10.1089/cmb.2012.0021.
27. Paton AW, Paton JC. 1998. Detection and characterization of Shiga toxigenic *Escherichia coli* by using multiplex PCR assays for stx_1, stx_2, *eaeA*, enterohemorrhagic *E. coli hlyA*, *rfbO111*, and *rfbO157*. J Clin Microbiol 36:598–602.
28. Scheutz F, Teel LD, Beutin L, Piérard D, Buvens G, Karch H, Mellmann A, Caprioli A, Tozzoli R, Morabito S, Strockbine 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. https://doi.org/10.1128/JCM.00860-12.
29. Lindsey RL, Garcia-Toledo L, Fasulo D, Gladney LM, Strockbine N. 2017. Multiplex polymerase chain reaction for identification of *Escherichia coli*, *Escherichia albertii* and *Escherichia fergusonii*. J Microbiol Methods 140:1–4. https://doi.org/10.1016/j.mimet.2017.06.005.
30. Ahmed R, Bopp C, Borczyk A, Kasatiya S. 1987. Phage-typing scheme for *Escherichia coli* O157:H7. Infect Dis 155:806–809. https://doi.org/10.1093/infdis/155.4.806.
31. 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. https://doi.org/10.1017/S0950268814000739.
32. International Organization for Standardization. 2012. ISO/TS 13136:2012. Microbiology of food and animal feed–real-time polymerase chain reaction (PCR)-based method for the detection of food-borne pathogens–horizontal method for the detection of Shiga toxin-producing *Escherichia coli* (STEC) and the determination of O157, O111, O26, O103 and O145 serogroups. International Organization for Standardization, Geneva, Switzerland.
33. Ashton PM, Nair S, Peters TM, Bale JA, Powell DG, Painset A, Tewolde R, Schaefer U, Jenkins C, Dallman TJ, de Pinna EM, Grant KA, Salmonella Whole Genome Sequencing Implementation Group. 2016. Identification of Salmonella for public health surveillance using whole genome sequencing. PeerJ 4:e1752. https://doi.org/10.7717/peerj.1752.
34. Chattaway MA, Schaefer U, Tewolde R, Dallman TJ, Jenkins C. 2017. Identification of *Escherichia coli* and *Shigella* species from whole-genome sequences. J Clin Microbiol 55:616–623. https://doi.org/10.1128/JCM.01790-16.
35. 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. https://doi.org/10.1128/JCM.01066-15.
36. Kohl TA, Diel R, Harmsen D, Rothgänger J, Walter KM, Merker M, Weniger T, Niemann S. 2014. Whole-genome-based Mycobacterium tuberculosis surveillance: a standardized, portable, and expandable approach. J Clin Microbiol 52:2479–2486. https://doi.org/10.1128/JCM.00567-14.
37. Day M, Doumith M, Jenkins C, Dallman TJ, Hopkins KL, Elson R, Godbole G, Woodford N. 2017. Antimicrobial resistance in Shiga toxin-producing *Escherichia coli* serogroups O157 and O26 isolated from human cases of diarrhoeal disease in England, 2015. J Antimicrob Chemother 72:145–152. https://doi.org/10.1093/jac/dkw371.
38. Kozyreva VK, Truong CL, Greninger AL, Crandall J, Mukhopadhyay R, Chaturvedi V. 2017. Validation and implementation of clinical laboratory improvements act–compliant whole-genome sequencing in the public health microbiology laboratory. J Clin Microbiol 55:2502–2520. https://doi.org/10.1128/JCM.00361-17.
39. Mellmann A, Andersen PS, Bletz S, Friedrich AW, Kohl TA, Lilje B, Niemann S, Prior K, Rossen JW, Harmsen D. 2017. High interlaboratory reproducibility and accuracy of next-generation-sequencing-based bacterial genotyping in a ring trial. J Clin Microbiol 55:908–913. https://doi.org/10.1128/JCM.02242-16.