PLOS ONE

RESEARCH ARTICLE

Nanopore sequencing for fast determination of plasmids, phages, virulence markers, and antimicrobial resistance genes in Shiga toxin-producing *Escherichia coli*

Narjol González-Escalona*, Marc A. Allard, Eric W. Brown, Shashi Sharma, Maria Hoffmann

Center for Food Safety and Applied Nutrition, Food and Drug Administration, College Park, MD, United States of America

* narjol.gonzalez-escalona@fda.hhs.gov

Citation: González-Escalona N, Allard MA, Brown EW, Sharma S, Hoffmann M (2019) Nanopore sequencing for fast determination of plasmids, phages, virulence markers, and antimicrobial resistance genes in Shiga toxin-producing *Escherichia coli*. PLoS ONE 14(7): e0220494. https://doi.org/10.1371/journal.pone.0220494

Editor: Chitrita DebRoy, The Pennsylvania State University, UNITED STATES

Received: May 2, 2019

Accepted: July 17, 2019

Published: July 30, 2019

Data Availability Statement: All relevant data are within the manuscript and its Supporting Information files.

Funding: This work was supported by the FDA Foods Science and Research Intramural Program and a grant to NGE from the MCMi Challenge Grants Program Proposal #2018-646. The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

Whole genome sequencing can provide essential public health information. However, it is now known that widely used short-read methods have the potential to miss some randomly-distributed segments of genomes. This can prevent phages, plasmids, and virulence factors from being detected or properly identified. Here, we compared assemblies of three complete Shiga toxin-producing *Escherichia coli* (STEC) O26:H11/H- genomes from two different sequence types (ST21 and 29), each acquired using the Nextera XT MiSeq, MinION nanopore-based sequencing, and Pacific Biosciences (PacBio) sequencing. Each closed genome consisted of a single chromosome, approximately 5.7 Mb for CFSAN027343, 5.6 Mb for CFSAN027346, and 5.4 MB for CFSAN027350. However, short-read whole genome sequencing (WGS) using Nextera XT MiSeq failed to identify some virulence genes in plasmids and on the chromosome, both of which were detected using the long-read platforms. Results from long-read MinION and PacBio allowed us to identify differences in plasmid content: a single 88 kb plasmid in CFSAN027343; a 157kb plasmid in CFSAN027350; and two plasmids in CFSAN027346 (one 95 Kb, one 72 Kb). These data enabled rapid characterization of the virulome, detection of antimicrobial genes, and composition/location of Stx phages. Taken together, positive correlations between the two long-read methods for determining plasmids, virulome, antimicrobial resistance genes, and phage composition support MinION sequencing as one accurate and economical option for closing STEC genomes and identifying specific virulence markers.

Introduction

Whole genome sequencing is an essential tool for characterizing and tracking pathogenic bacteria that may have contaminated the food supply as well as for identifying whether those bacteria carry virulence factors that could pose serious threats to public health [1,2]. Closed

Competing interests: The authors have declared that no competing interests exist.

bacterial genomes provide: 1) high-quality reference material that supports pathogen source tracking during a foodborne outbreak investigation, 2) important clues about the long-term evolution of enteric pathogens, 3) key insights into mechanisms and transmission of mobile elements conferring antimicrobial resistance, and 4) critical information about the potential contribution of DNA modifications on pathogenesis [2–9]. These details are particularly important for understanding hemorrhagic pathogens such as Shiga toxin-producing *E. coli* (STEC) O26:H11/H- strains, which cause significant human morbidity and mortality worldwide [10–13]. The genomes of O26:H11/H- are very complex, containing many virulence genes, insertion sequences, phages, and plasmids; consequently, strains of the same lineage can possess significantly different content [7,9,10],[11,14–16]. Missing the presence of some of these elements during an investigation can have large impacts on human health [e.g. *hlyA* gene in enterohemorrhagic *E. coli* (EHECs)]. Thus, it really matters that we have reproducible WGS systems for fully capturing and sequencing these elements.

Long-read sequencing platforms afford one solution to this challenge. Some systems such as Pacific Biosciences (PacBio) Sequencers *RSII* or Sequel (https://www.pacb.com/products-and-services/pacbio-systems/), use single-molecule real-time (SMRT) sequencing technology that allow for real-time observation of DNA synthesis through zero-mode waveguides (ZMWs) and phospho-linked nucleotides [17,18]. While comprehensive in their ability to capture entire genomes, extraneous elements included, these systems often require significant investments in machinery, space, and laboratory expertise, all of which may be obstacles to routine use. These systems also require significant quantities of DNA (*i.e.*, 5 μg), require a more substantial preparatory time (i.e., 8 hr DNA sequencing library protocols), and produce average read lengths of about 11Kb, although reads of greater than 50 kb were possible in the laboratory at the time of this study.

Alternative sequencing platforms based on nanopore technology may be able to provide high-quality libraries from long reads and produce closed bacterial genomes, while also offering several other advantages in portability and affordability. These systems are much less expensive, take little laboratory space, and can even be taken into the field for on-site sequencing. The MinION nanopore (Oxford Nanopore, Oxford, UK) system, as one example, comprises a palm-sized unit able to detect changes in ionic current when DNA or RNA passes through the nanopores, whereupon those changes are translated into base calls. Researchers have already used nanopore sequencing on plants, yeasts, viruses, and to perform *de novo* bacterial assembly [19–21]. Other applications have included rapid identification of viral pathogens [22,23], metagenomics [23–25], detection of antimicrobial resistance genes [26,27], comparative RNA expression levels in diverse cells [28], and detecting DNA methylation patterns [29]. Since the initial MinION Access Programme (MAP), this technology has been refined several times [30]. At least three nanopore-based sequencing tools are currently in production as of autumn 2018 [19,21,30,31].

As MinION does not require a size selection step prior to sequencing, it can obtain longer reads than other platforms [19,22,30–33]. Moreover, the system also allows for "quasi" real-time sequencing approaches [34]. The accuracy of nanopore sequencing is reported to be 90% in general, with some researchers reporting 99.96% accuracy only after read polishing (http://simpsonlab.github.io/2016/08/23/R9).

Two of the main subgroups within O26:H11/H- are the enteropathogenic (EPEC), which carry the locus of enterocyte effacement (LEE) causing mild diarrhea, and enterohemorrhagic (EHEC), which carry either *stx*1 and/or *stx*2 genes in addition to the LEE, and are associated with more severe illnesses, such as hemorrhagic colitis (HC) and hemolytic uremia syndrome (HUS) [35,36]. A new clone of O26:H11/H- belonging to sequence type 29 (ST29) and that had a specific virulome ($stx2_a$+, *eae*+, plasmid gene profile *ehxA*+, *etpD*+) has been recently

found distributed all over Europe [10]. Serotype O26:H11/H- can be divided, based on MLST, into several STs, with ST21 and ST29 associated with disease in humans [7]. We have selected three STEC O26:H11 strains with very complex genomes belonging to different STs (21 and 29), and isolated in different years, places and sources (clinical and environmental). This makes them ideal strains to compare the three WGS capabilities.

In this study, we compared the sequencing capabilities of the MinION with that of the PacBio RS II and MiSeq by sequencing 3 STEC O26:H11 strains [10] using MinION and a genome library prepared using two different library kits (*i.e.*, 1D ligation kit and rapid sequencing kit). We then compared the results obtained by this sequencing platform (MinION). We also tested two different assembly pipelines for the *de novo* assembly of these genomes. Finally, we compared the results of each technology to assess their capacity for detecting virulence and antimicrobial genes, Shiga toxin phages, plasmid presence, genome synteny, and phylogenetic analysis.

Results

The 3 STEC strains employed here are listed in Table 1, along with their sources and confirmed STs. These genomes were sequenced on each of the sequencing systems, and three main differences were documented among those systems: number of contigs, ease of assembly, and detection of important genomic markers.

MiSeq sequencing and assembly

These STECs were sequenced using using Nextera XT MiSeq (Illumina) and assembled *de novo*. Their genomes assembled into 234 to 321 contigs (Table 2), which is within the expected range for *de novo* assembled STECs. Results from other drafts STEC genomes are typically around >250 contigs (based on our own and other researcher's observations; data available from NCBI). *In silico* MLST identified the samples correctly as belonging to ST21 (CFSAN027343 and CFSAN027346), and ST29 (CFSAN027350). *in silico* serotyping confirmed that these strains were all serotype O26:H11.

De novo assembly of the 3 genomes resulted in assemblies that ranged from 5.2 (CFSAN027350) to 5.4 Mb (the other two genomes) (Table 2). As stated earlier, it is notable that *de novo* assemblies of short-read sequencing technology produced an elevated number of contigs resulting in fractioned assemblies.

MinION sequencing and DNA sequencing library approaches

We sequenced the same three STECs using the nanopore-based MinION device. Two strains were sequenced in triplicate (CFSAN027350 and CFSAN027346); CFSAN027343 was sequenced twice. We used two different DNA library approaches: the 1D ligation kit (SQK-LSK108, expected to produce more output and longer reads) and the rapid sequencing

Table 1. Summary of the characteristics of the 3 STEC O26:H11 strains sequenced in this study.

strain	CFSAN No.	Serotype[a]	ST[b]	Date	location	source
99.085	CFSAN027343	O26:H11	21	1999	Argentina	clinical
99.1773	CFSAN027346	O26:H11	21	1999	USA	clinical
12.1843	CFSAN027350	O26:H11	29	2012	USA	environmental

[a]These results were confirmed with *in silico* serotyping.

[b]Determined by *in silico* MLST.

https://doi.org/10.1371/journal.pone.0220494.t001

Table 2. MiSeq assembly statistics for the 3 O26:H11 STECs.

CFSAN No.	Contigs MiSeq	No. of reads	Q>30 Reads size	Total bases (bp)	N50	Total genome (bp)	Average coverage X
CFSAN027346	274	2.14E+06	194	4.17E+08	100,594	5.42E+06	77
CFSAN027343	321	2.56E+06	188	4.81E+08	93,611	5.41E+06	89
CFSAN027350	234	2.45E+06	221	5.42E+08	99,163	5.28E+06	103

https://doi.org/10.1371/journal.pone.0220494.t002

kit (SQK-RAD002, expected to produce lower output and smaller reads, but with a much simpler and straightforward procedure for preparing the DNA library).

Table 3 confirms that the 1D ligation kit produced larger outputs than the rapid sequencing kit. Even though the DNA extraction method was the same for each strain, the sequencing output was different, varying from 1 Gb (CFSAN027350a) to 8 Gb (CFSAN027343a). DNA libraries output also varied between replicates (with a higher percentage of reads > 5 kb in the case of CFSAN027346b with 355,653 reads above 5 kb long compared to the same replicate library CFSAN027346a with 139,883 reads above 5 kb (results not shown). As expected, sequencing runs prepared with the rapid kit produced lower output, with an average of 0.71 Gb (Table 3).

Nevertheless, after running our CANU assembly, we were able to produce a closed bacterial genome for each strain, including the chromosome and plasmid(s) (Table 4). Although there were still variations in chromosome sizes among the replicates, we found overall agreement that CFSAN027343 and CFSAN027350 each contained a chromosome of ~ 5,6 Mb with a single plasmid while CFSAN027346 contained two plasmids. We will describe this plasmid in more detail later in this section.

PacBio sequencing

We next sequenced our three STECs using the PacBio *RSII* system with a 20-kb insert library protocol. After the library protocol was completed, we sequenced it on three SMRT cells (Table 5). Although the same preparation was used for each isolate, slight differences were noted in the output of raw data among different SMRT cells, (Table 5). The total output of the three SMRT cells for CFSAN027343, CFSAN027346, and CFSAN027350 was 2.3, 2.84. and 2.9 Gb, respectively. The small discrepancy is common since the SMRT cells were loaded with a different binding complex for each strain. The average read of insert lengths for each

Table 3. STECs MinION sequencing output statistics by replicate and library kit.

Run[1]	Total reads	total output (GB)	Average coverage (X)
[2]CFSAN027343a	4,027,578	8.21	1127
[2]CFSAN027346a	692,153	3.33	585
[2]CFSAN027346b	611,279	6.82	1199
[2]CFSAN027350a	267,065	1.08	189
[2]CFSAN027350b	1,369,072	5.23	920
[3]CFSAN027343	173,937	0.93	163
[3]CFSAN027346	155,628	0.90	158
[3]CFSAN027350	131,153	0.31	54

[1]Runs: a) first, and b) second run.

[2]1D Genomic DNA by ligation (SQK-LSK108)

[3]rapid sequencing kit (SQK-RAD002)

https://doi.org/10.1371/journal.pone.0220494.t003

Table 4. Assembly statistics MinION.

Sample[1]	Chromosome(s) contig(s)	plasmids	Chromosome size (bp)	plasmid(s) size(s) (bp)
[2]CFSAN027343a	1	1	5,688,712	88,561
[3]CFSAN027343	1	1	5,688,145	88,702
[2]CFSAN027346a	1	2	5,588,947	1 (95,599); 2 (72,940)
[2]CFSAN027346b	1	2	5,592,589	1 (95,821); 2 (72,972)
[3]CFSAN027346	1	2	5,592,692	1 (95,696); 2 (72,950)
[2]CFSAN027350a	1	1	5,422,984	157,276
[2]CFSAN027350b	3	1	5,448,646	157,340
[3]CFSAN027350	15	1	5,451,905	157,300

[1]Runs: a) first, and b) second run.

[2]1D Genomic DNA by ligation (SQK-LSK108)

[3]rapid sequencing kit (SQK-RAD002)

https://doi.org/10.1371/journal.pone.0220494.t004

individual strain CFSAN027343, CFSAN027346, and CFSAN027350 was 10,011 bp, 10,804 bp, and 11,632 bp, respectively.

To compare assembly performance from Pacbio data, we carried out *de novo* assembly for every strain with each SMRT cell sequenced. Only for CFSAN027350 the data from three SMRT cells were combined. The overall consensus concordance for the chromosome was between 99.91% and 99.99% and for the plasmids, between 99.95% and 99.99%. All three *de novo* assemblies for CFSAN027346 and CFSAN027350 generated a single contig for the chromosome and the plasmid(s) (Table 6). The three assemblies made using one SMRT cell for strain CFSAN027343 produced a single contig for the plasmid, but 2–3 contigs for the chromosome, due to the presence of a larger repeat (Table 6). To achieve a single contig for the chromosome, we had to combine the three SMRT cells and set the filter for the minimum subread length to 5000. Using these settings, the chromosome was able to be closed for CFSAN027343 with a coverage of 400X. After manual closure, a Quiver consensus algorithm was run for each consensus thereby achieving a consensus concordance of 100% with an average coverage from 130X to 170X. It is important to note that the sizes of the chromosome identified for CFSAN027346 and CFSAN027350, and the sizes of the plasmids found after three different SMRT cells runs for each strain were almost identical (variation $< 0.0003\%$).

Table 5. PacBio sequencing output for each SMRT cell.

Strains	SMRT Cell#	Total Reads	Total Output (Gb)	Average Read Length of Insert
CFSAN027343	1	79,102	0.79	9,965
	2	77,649	0.8	10,326
	3	75,856	0.74	9,743
CFSAN027346	1	91,744	1.03	11,252
	2	82,672	0.88	10,677
	3	87,886	0.93	10,484
CFSAN027350	1	81,504	0.95	11,715
	2	82,543	0.96	11,641
	3	87,866	1.01	11,541

https://doi.org/10.1371/journal.pone.0220494.t005

Table 6. Assembly statistics per SMRT cell# for PacBio data using HGAP3.0 and Quiver.

Strains[a]	Chromosome contig(s)[b]	Chromosome size (bp)[c]	Plasmids contig(s)[d]	Plasmid(s) size(s) (bp)[e]	Coverage (X)[f]
CFSAN027343					
SMRT Cell # 1	3	5,351,371; 281,039; 81,920	1	88,847	134
SMRT Cell # 2	2	5,525,151; 164,081	1	88,848	134
SMRT Cell # 3	2	5,525,031; 164,076	1	88,847	129
CFSAN027346					
SMRT Cell # 1	1	5,592,579	2	1 (96,016), 2 (73,152)	147
SMRT Cell # 2	1	5,592,570	2	1 (96,016), 2 (73,152)	156
SMRT Cell # 3	1	5,592,582	2	1 (96,017), 2 (73,152)	153
CFSAN027350					
SMRT Cell # 1	1	5,436,071	1	157,534	170
SMRT Cell # 2	1	5,436,072	1	157,534	166
SMRT Cell # 3	1	5,436,082	1	157,535	176

[a]The statistic is listed for each SMRT cell# per isolate.

[b]The # represents the assembled contigs that belong to the chromosome.

[c]The size (bp) for each assembled contig that belongs to the chromosome.

[d]The # represents the assembled contigs that belong to the plasmid(s).

[e]The size (bp) for each assembled contig that belongs to the plasmid(s).

[f]The number represents the mean coverage for the assembly for each SMRT cell per isolate.

https://doi.org/10.1371/journal.pone.0220494.t006

Detection of virulence genes

The output assemblies from our MinION/CANU assembly were used for *in silico* detection of 95 described virulence genes for *E. coli* [37], which include genes reported from all pathotypes of pathogenic *E. coli*, with particular focus on EHECs. To test our hypothesis that Nextera XT MiSeq library sequencing was subpar for obtaining a complete representation of the STEC genomes (chromosome and plasmids), we compared results obtained using MinION assemblies against those obtained from Nextera XT MiSeq assemblies. We confirmed those results using the PacBio assemblies as reference.

The 3 O26:H11 STEC strains were positive for 18 of the 95 virulence genes tested (Table 7). Among those genes were: *astA*, *cif*, *eae*, *ehxA* (plasmid), *espA*, *espB*, *espF*, *espJ*, *espP* (plasmid), *gad*, *iha*, *iss*, *lpfA*, *nleA*, *nleB*, *nleC*, *tir*, and *toxB* (plasmid) genes. Other genes were sporadic and strain dependent: *tccP* gene was present in CFSAN027346 and CFSAN027350, *efa1 and katP* (plasmid) genes were only present in CFSAN027343 and CFSAN027346. CFSAN027350 was the only strain found to contain *espI* and *stx2a*. CFSAN027343 and CFSAN027346 carried a different Shiga toxin phage that instead contained the *stx1a* variant.

MinION assemblies were congruent and showed the presence/absence of the same genes for each strain (Table 7). We found both nanopore DNA libraries (ligation and rapid) provided sufficient data to assemble and include all virulence genes present, confirming that nanopore technology can provide fast determination of virulence potential by detecting specific virulence genes. Detection of the same virulence genes was also observed with the PacBio assemblies for each strain. However, several virulence genes (*i.e.*, *toxB*, *tccP*, *iha*, and *astA*) were not detected in some of the genome assemblies obtained with Nextera XT MiSeq, pointing to the corresponding library prep as having been responsible for loss of some of these segments or problems with the assembly of those gene regions.

Table 7. Virulence genes present in the O26:H11 STEC MinION assemblies by *in silico* analysis. Plasmid borne genes were: *espP*, *toxB*, *katP*, and *ehxA*.

Assemblies[1]	*espI*	*stx2a*[4]	*stx1a*[5]	*tccP*	*efa1*	*katP*
[2]CFSAN027343a	-	-	+	-	+	+
[3]CFSAN027343	-	-	+	-	+	+
[2]CFSAN027346a	-	-	+	+	+	+
[2]CFSAN027346b	-	-	+	+	+	+
[3]CFSAN027346	-	-	+	+	+	+
[2]CFSAN027350a	+	+	-	+	-	-
[2]CFSAN027350b	+	+	-	+	-	-
[3]CFSAN027350	+	+	-	+	-	-

All assemblies were positive for *astA*, *cif*, *eae*, *ehxA* (plasmid), *espA*, *espB*, *espF*, *espJ*, *espP* (plasmid), *gad*, *iha*, *iss*, *lpfA*, *nleA*, *nleB*, *nleC*, *tir*, and *toxB* (plasmid) genes.

[1]Runs: a) first, and b) second run.

[2]1D Genomic DNA by ligation (SQK-LSK108).

[3]rapid sequencing kit (SQK-RAD002).

[4,5]shiga toxin genes and their variants.

https://doi.org/10.1371/journal.pone.0220494.t007

stx phage identification and location

We assessed whether MinION data could precisely locate where Shiga toxin phages and major pathogenicity islands were located on the genomes, comparing those results with PacBio data–MiSeq was unable to reconstruct individual phages from these strains.

For the MinION-assembled chromosome CFSAN027343, we detected 20 prophage regions using Phaster [38], of which: 14 regions were intact, 3 regions were incomplete, and 3 regions were questionable (a sample Phaster result is shown in S3 Fig). The *stx* carrying phage was 57.6 Kb located in the *torS-torT* intergenic region gene, different from what was previously known for *stx* phage insertions (Table 8). The PacBio sequence gave the same number of prophage regions for this genome. Also the *stx1* phage in the PacBio assembly was of the same size and in the same location as observed for the MinION assembly.

In the MinION-assembled chromosome in CFSAN027346, Phaster detected 17 prophage regions: 15 regions were intact, 1 region was incomplete, and 1 region was questionable. PacBio sequence for the same genome was similar, although only 16 prophages were identified (14 intact, 1 incomplete and 1 questionable). Surprisingly, akin to PacBio, that particular *stx* phage was at the classic *wrbA* gene insertion site. Nonetheless, both MinION and PacBio found the size of the phage to be 69.3 kb (Table 8).

Table 8. Identification of chromosomal insertion sites for stx phages in the 3 STEC O26:H11 MinION genomes, their *stx* gene type, regions and coordinates in the genome, and their *stx* phage sizes.

	Insertion of Stx phage in[a]								
Strains	***wrbA***	***yehV***	***yecE***	***sbcB***	***torS-torT intergenic region***	***mlaA—ypdK intergenic region***	**Region phaster- stx phage (coordinates genome)**	***stx* phage size (kb)**	***stx type***
CFSAN027343	-	-	-	-	+	-	14 (3518081–3575683)	57.6	1a
CFSAN027346	+	-	-	-	-	-	8 (2329980–2401033)	69.3	1a
CFSAN027350	-	-	-	-	-	+	14 (3749826–3852495)	102.6	2a
11368 (NC_013361.1)	+	-	-	-	-	-	9 (2347644–2418764)	69.3	1a

[a] All chromosomes started at the *dnaA* gene.

https://doi.org/10.1371/journal.pone.0220494.t008

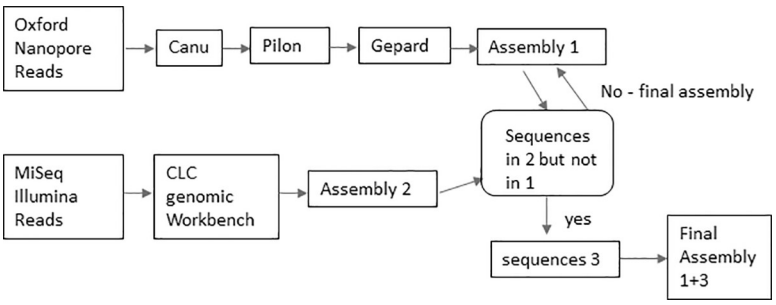

Fig 1. Schematic representation of the analysis pipeline used in this study for assembly and polishing of the MinION sequencing output.

https://doi.org/10.1371/journal.pone.0220494.g001

In the MinION-assembled chromosome in CFSAN027350, Phaster detected 17 prophage regions: 13 regions were intact, 1 region was incomplete, and 3 regions were questionable. PacBio sequence for the same genome identified 16 prophages (11 intact, 5 incomplete; none were questionable). Again, the phage carrying *stx* was located at an unusual insertion site, *mlaA—ypdK* intergenic region, and this *stx* phage was the biggest among the 3 strains– 102.6 kb (Table 8).

Detection of antimicrobial genes

In silico detection of antimicrobial resistance (AMR) genes using the output assemblies from our MinION pipeline found that only CFSAN027346 carried antimicrobial resistance genes, specifically: *aph*(3")-Ib, *aph*(6)-Id, *bla*TEM-1B, *sul2*, *tet*B, and *dfrA*. All these genes were contained within a smaller plasmid (73 kb) (Table 4 and Table 6) (S1 Fig). Using the assembly generated by the Nextera XT MiSeq data, *in silico* analysis showed a similar result but missed the *sul2* gene.

cgMLST SNP analysis of O26:H11 genomes–MiSeq, MinION, and PacBio

For this section, *de novo* assemblies were produced for the nanopore data using our analysis pipeline described in Fig 1 which includes assembly and Nextera XT polishing of the MinION/CANU assemblies. The phylogenetic relationships among *E. coli* O26:H11/H- strains were determined by a cgMLST SNP analysis shown in Fig 2. This cgMLST was the same as previously published [9]. The genome of *E. coli* strain 11368 (NC_013361.1) was used as a reference and has 4,554 genes. The resultant NJ SNP tree showed that the O26 genomes analyzed were highly diverse and polyphyletic and that the assemblies for these 3 strains (*i.e.*, MinION polished assemblies) clustered with the assemblies of the genomes for the same strain generated by the other two technologies that are known to be more accurate (*i.e.*, the MiSeq and PacBio data) (Fig 2A). However, the assemblies generated by our pipeline were located in longer branches in the same cluster and showed that they still have many errors that differentiated them from their counterparts of high-quality genomes generated by MiSeq or PacBio by 128–203 SNPs for CFSAN027343 (Fig 2C and S2A Fig) and 92–175 SNPs for CFSAN027346 (Fig 2C and S2B Fig).

Synteny comparisons of MinION and PacBio assemblies

A comparison of genome synteny between the assemblies produced by MinION/CANU *vs*. PacBio was performed using Mauve aligner [39]. Overall MinION-generated chromosomes for the 3 strains have the same synteny to the ones generated by PacBio (Fig 3). (Fig 3).

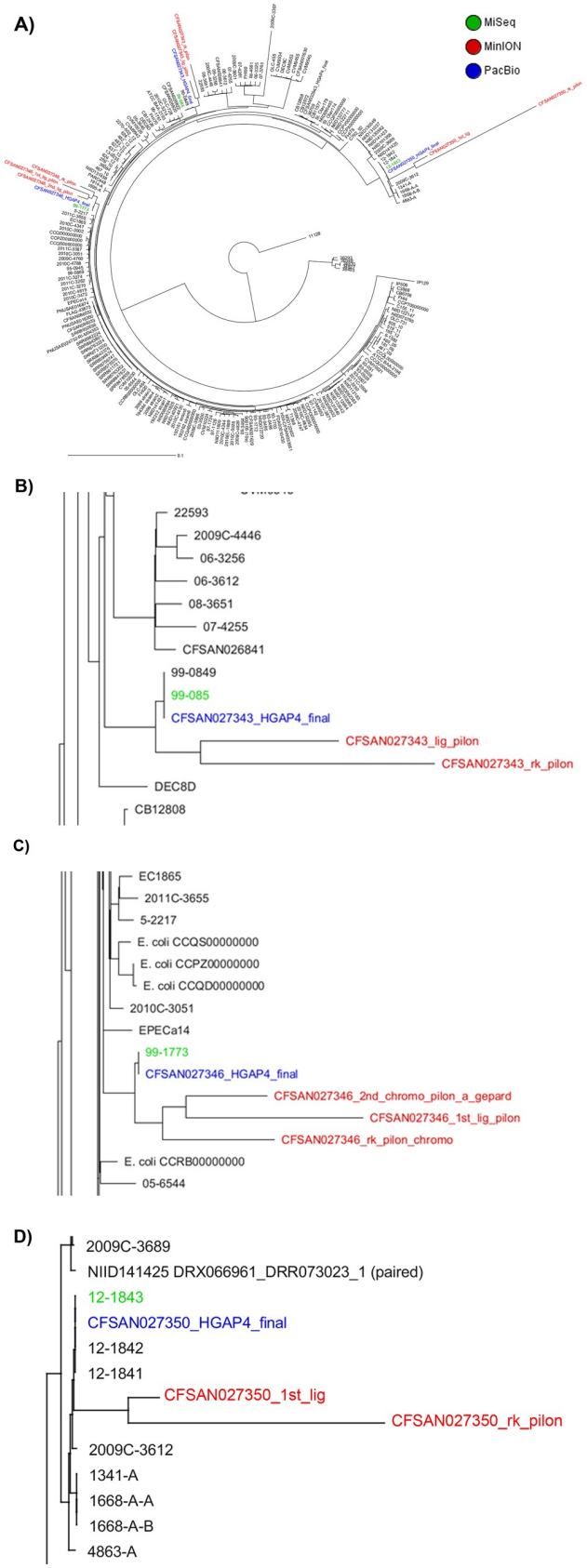
A)
MiSeq
MinION
PacBio
B)
22593
2009C-4446
06-3256
06-3612
08-3651
07-4255
CFSAN026841
99-0849
99-085
CFSAN027343_HGAP4_final
CFSAN027343_lig_pilon
CFSAN027343_rk_pilon
DEC8D
CB12808
C)
EC1865
2011C-3655
5-2217
E. coli CCQS00000000
E. coli CCPZ00000000
E. coli CCQD00000000
2010C-3051
EPECa14
99-1773
CFSAN027346_HGAP4_final
CFSAN027346_2nd_chromo_pilon_a_gepard
CFSAN027346_1st_lig_pilon
CFSAN027346_rk_pilon_chromo
E. coli CCRB00000000
05-6544
D)
2009C-3689
NIID141425 DRX066961_DRR073023_1 (paired)
12-1843
CFSAN027350_HGAP4_final
12-1842
12-1841
CFSAN027350_1st_lig
CFSAN027350_rk_pilon
2009C-3612
1341-A
1668-A-A
1668-A-B
4863-A

Fig 2. Phylogenetic analysis of the O26:H11/H- *E. coli* strains sequenced in this study by MiSeq, MinION, and PacBio and 195 genomes that are available at GenBank by cgMLST analysis. The SNPs were extracted from the core loci (1303) and the SNP matrix (5089 SNPs) was used to determine the genetic relationships among the strains. The evolutionary history was inferred by using Neighbor-Joining (NJ) tree built using the genetic distance and showing the existence of high diversity and that O26:H11 strains were polyphyletic. The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. The tree was rooted to *E. coli* O111:H- strain 11128 (NC_013364). A) The genomes generated by any of the 3 technologies still clustered together by the cgMLST analysis. Snapshot of the clusters formed by the genomes generated by the 3 technologies for B) CFSAN27343, C) CFSAN027346, and D) CFSAN027350 strains, respectively. The names of the strains can be discerned in S5 Fig.

https://doi.org/10.1371/journal.pone.0220494.g002

Novel plasmid in O26:H11 (155 kb)

The virulence plasmids of CFSAN027343 and CFSAN027346 were 88 kb and 96 kb, respectively. Interestingly, the virulence plasmid for CFSAN027350 was much larger (157 kb). Both pCFSAN027343 and pCFSAN027346, however, carried these 4 virulence genes: *ehxA*, *espP*, *toxB*, and *katP*, while pCFSAN027350 carried only *ehxA*, *espP*, and *toxB*. Comparative analyses of the 3 plasmids showed that pCFSAN027350 was very different from the other 2 plasmids (Fig 4), possessing extensive unique regions. Based on these data, pCFSAN027350 contains 214 annotated ORFs with multiple transposons and insertion elements and has a G + C content of 48.4% (S2 Table).

Discussion

Based on MLST analyses, strains of STEC O26:H11/H- can be separated into several STs; of these, ST21 and ST29 have been associated with disease in humans [7]. Therefore, it is especially important to be able to rapidly characterize all detected STECs O26:H11/H- (both chromosome and plasmid(s), if present) to determine whether those clones or strains circulating in particular locations are likely to be threats to human health. Here, we tested the use of such a technique (*i.e.*, Nanopore sequencing) to obtain the complete genomes of 3 unrelated STEC O26:H11 strains followed by subsequent comparisons to those obtained by PacBio and Nextera XT MiSeq WGS platforms. The 3 strains belonged to the EHEC O26 group in accordance with the O26:H11/H- cgMLST scheme reported previously by Gonzalez-Escalona et al. in 2016 [9].

Fig 3. Comparison of the synteny mapping of chromosomes generated by either MinION or PacBio, using the Sanger generated genome for STEC O26:H11 strain 11368 (AP010953) as reference with MAUVE. Each chromosome sequence is laid out in a horizontal track. Matching colors indicate homologous segments and are connected across genomes. Respective scales show the sequence coordinates in base pairs. A colored similarity plot is shown for each genome, the height of which is proportional to the level of sequence identity in that region. Only strain CFSAN027343 synteny is shown for illustration purposes (The other two strains (CFSAN027346 and CFSAN027350) synteny can be found in S4 Fig).

https://doi.org/10.1371/journal.pone.0220494.g003

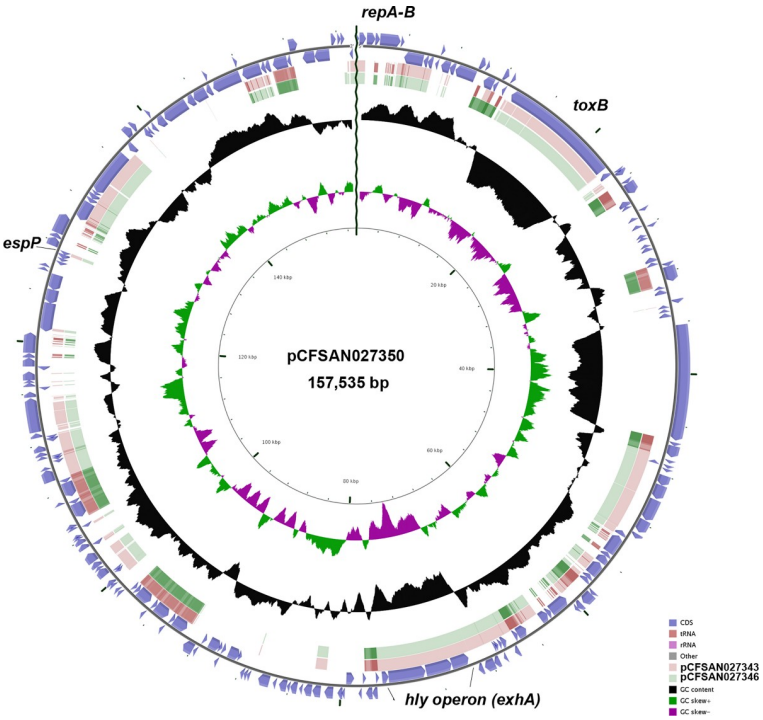

Fig 4. Circular map of virulence plasmid pCFSAN027350 compared to the other two virulence plasmids (pCFSAN027343 and pCFSAN027346), generated using CGView [63]. Blue block arrows in the outer circle denote coding regions in the plasmid, indicating the ORF transcription direction. G+C content is shown in the middle circle and the deviation from average G+C content (47.71%) is displayed in the innermost circle. BLAST comparisons with the other two EHEC plasmids are shown in light red (pCFSAN027343) and green (pCFSAN027346).

https://doi.org/10.1371/journal.pone.0220494.g004

We conducted Nanopore testing using a MinION device connected to a laptop using two different DNA sequencing kits (the 1D ligation kit and the rapid sequencing kit) in the study. Resultant genomes were able to be closed using both DNA sequencing kits deployed here save for CFSAN027350. For that strain we could not generate a complete closed genome using the rapid kit—the chromosome was contained in 15 contigs. Nonetheless, the virulence plasmid for that strain/run was contained in a single contig. One explanation for this discrepancy could be that the WGS output was very low at <0.14 Gb for reads above 5,000 bp (*i.e.*, 21,759 total reads). Also, the minimum amount of reads that were necessary for closing the STEC genome was around 58,000 reads above 5,000 bp for the rapid kit. The number of reads generated by the 1D ligation kit was around 5 times higher than that for the rapid kit, which allowed us to estimate that, in principal, we could run up to 5 different strains per flow cell thus reducing the cost per genome to ~200 USD. This assumption is reinforced by the newer released RAD004 kit, which produces even higher and better-quality output than its predecessor (unpublished observations).

In the last 4 years, Pacbio sequencing, combined with *de novo* assembly, has been an attractive method for closing bacterial chromosomes and the plasmid(s) that they may carry. Indeed, in this study, we verified that by using PacBio technology with P6-C4 chemistry and *de novo* assembly, the three *E. coli* genomes including their plasmids studied here were able to be completely sequenced and closed. The three assemblies obtained for each isolate using the raw sequence data from each SMRT cell run, respectively, showed high consistency and reproducibility with a coverage of above 100X and an average read length of 10 to 12 kb. Moreover, all chromosomes and plasmids could be completely closed with the sequence data from one

SMRT cell except for the single chromosomal sequence run from CFSAN027343. Fragmented assemblies can occur when the repetitive region in some bacteria is longer than the read lengths. Pathogenic *E. coli*, in fact, are known for carrying repetitive regions within the chromosome and read lengths obtained from the PacBio sequencer, at times, are not long enough to overcome this challenge. Albeit, Pacbio is a stand alone instrument that does not require finishing or polishing with additional WGS data from another source.

Using MinION-generated genomes, polished with Nextera XT MiSeq reads, allowed us to accurately and efficiently determine the virulence genes present in each strain, as well as which genes were located on plasmid(s) or on the chromosome (see Table 7). We also were able to determine the presence of AMR in one of the strains (CFSAN027346) and that was located on an extra plasmid carried solely by that strain (S1 Fig). The other two STEC strains carried a single virulence plasmid albeit with different sizes and gene compositions. It was also noteworthy that none of these strains carried the *etpD* gene (coding for a Type II secretion protein) in the plasmid, indicative of the presence of the European clone [9,10,40,41]. From the nanopore data analyzed here, the virulence profile for each strain was determined as follows: CFSAN027343 (ST21, stx^{1a}+, *eae-beta1*+, plasmid gene profile *ehxA*+, *espP*+, *katP*, and *toxB* +), CFSAN027346 (ST21, stx^{1a}+, *eae-beta1*+, plasmid gene profile *ehxA*+, *espP*+, *katP*, and *toxB*+), and CFSAN027350 (ST29, stx^{2a}+, *eae-beta1*+, plasmid gene profile *ehxA*+, *espP*+, and *toxB*+).

Besides differences in their Shiga toxin phage content, we detected other virulence genes unique to some of the 3 STECs analyzed here. Among them was the *tccP* gene, coding for an effector protein (Tir-cytoskeleton coupling protein), found commonly in O157:H7 EHEC strains [42,43] and which was present in strains CFSAN027346 and CFSAN027350. The *tccP* gene plays a direct role during EHEC infection inducing actin polymerization by coupling Tir to the actin cytoskeleton [44]. The presence of this gene, together with *espJ*, *stx1*a or 2a, intimin, and Tir, makes these strains highly pathogenic. Another important illustration of the mosaic of virulence factors found within STEC O26:H11 strains is that two important genes for intestinal colonization (*efa1/lifA*—a protein with putative glycosyltransferase activity that has an important role in intestinal colonization), *and katP- (*a catalase-peroxidase, which might help EHEC O157:H7 to colonize host intestines by reducing oxidative stress), were present in both CFSAN027343 and CFSAN027346 [45,46].

We also sequenced the same 3 strains using PacBio, which is considered the gold standard for closing bacterial genomes [18,47,48]. These genomes were compared to the ones generated by nanopore sequencing. Overall, there was fair agreement with the sequences in both the chromosome(s) and plasmid(s) generated from the 3 STEC strains. This was evidenced by the extremely high synteny as well as virulence genes profiles, AMR genes, and overall plasmid content. Despite this extraordinary congruence, some discrepancies were observed for CFSAN027343 where there was a fragment (~ 1.5 Mb) that appears to be reversed in direction on the chromosome between the two, influencing the resultant calling of the *stx* phage and with MinION *stx* phage being smaller (57.6 kb) than the PacBio stx-phage (75.6 kb).

As observed by other authors, sequencing accuracy continued to be an issue for the genomes generated by nanopore sequencing here. That is, PacBio maintained 99% accuracy (1 SNP observed by cgMLST) while MinION genomes revealed expectedly lower accuracy (92–203 SNPs by cgMLST) (S2 Fig). Without Nextera XT MiSeq polishing, we are unable to accurately place the MinION assemblies -using this cgMLST scheme- into the phylogenetic tree as clustering relies critically on sequence calling using complete ORFs. One of the main drawbacks of MinION nanopore sequencing is that the assemblies contained numerous artifactual indels that will readily translate into incorrect allele calls. This will, in turn, affect the correct clustering of those isolates in the phylogenetic tree. Continued improvements in nanopore-

based sequencing chemistry (*e.g.* RAD004) as well as development of alternative base-calling algorithms such as Scrappie–an open-source transducer neural network [49] may further mitigate this problem. This latter development is designed to aid in the correction of longer homopolymeric runs, one of the main challenges of the nanopore platform. On the other hand, it is equally important to note that PacBio also has continued to mitigate caveats including to time of labor, and read length with recent advances in Sequel chemistry (SMRTbell Express Template Prep Kit, Sequel Sequencing and Binding Kit 3.0 and SMRT Cell 1M v3).

By using MinION sequencing we were able to easily identify the location and composition of the shiga toxin carrying phages in the STEC O26:H11 strains. Contrary to what has been observed for EHEC O157:H7 and non-O157:H7 STECs, the location on the chromosome of the stx phages [50] were in different regions of the chromomere for each strain in this study. According to Bonanno et al. (2015), nine Stx phage insertion sites have been described in STEC strains genes: *wrbA*, *yehV*, *yecE*, *sbcB*, *Z2577*, *ssrA*, *prfC*, *argW*, *torS-torT* intergenic region. Delannoy et al. (2017) also described the presence of a Stx phage at a different insertion site *yciD* for O26:H11/H- strains [51]. For the prototypic Sanger sequence STEC O26:H11 strain 11368, the Stx phage is located at the *wrbA* gene, a very commonly observed site for *stx* phage insertions in STECs [50]. The Stx phage in these 3 strains (sequenced in this study) were located known locations on the chromosome (*torS-torT* intergenic region–CFSAN027343, *wrbA*- CFSAN027346) except for CFSAN027350 which was located at a novel location in the intergenic region between genes *mlaA* and *ypdK*, highlighting once more the high diversity among STECs O26:H11/H- strains.

Finally, with both MinION and PacBio, we were able to identify a large virulence plasmid in strain CFSAN027350 (~ 157 kb) (Fig 4). The presence of a large virulence plasmid in O26:H11 strains is not uncommon as was observed for strain H30 (pO26-Vir—168 kb) [14], which contained 5 additional plasmids. As was observed with pO26-Vir, pCFSAN027350 showed a mosaic structure, with many fragments of the plasmid matching other plasmids available at GenBank and contained multiple transposons and insertion sequences, evidence of the mobility of some of the regions (e.g. *toxB* gene was surrounded by IS elements at both ends of the gene). Nevertheless, this plasmid was highly similar (99%) to another plasmid reported for a clinical strain of O26:H11, albeit of larger size (~181 kb, strain 2013C-3277 plasmid unnamed3, CP027334.1).

Overall, the high degree of correlation we found between these two long-read methods for determining plasmids, virulome, antimicrobial resistance genes, and phage composition in STEC O26 strongly indicates that the MinION sequencing technology is an excellent solution for rapidly determining STEC O26 closed genomes and performing comprehensive analysis of their genomic markers. The MinION devices are robust enough to be used to monitor viruses in remote areas [21], yet can provide most of the applications provided on the PacBio platform (methylation patterns, metagenomics, finding location of pathogenicity islands, resolving plasmids, among other applications) [24–26,29,31,32]. To be able to rapidly provide an assessment of possible virulence, antimicrobial resistance potential and disease risk posed by any STEC using this nanopore sequencing technology, is of paramount importance for the protection of the public health and for the tracking of existent and new clones in any geographical region.

Materials and methods

Bacterial strains and media

E. coli strains CFSAN027343, CFSAN027346, and CFSAN027350 (Table 1) were purchased from the *E. coli* Reference Center (Pennsylvania State University, University Park, PA). These strains were revived in Tryptic Soy Broth and grown overnight at 37˚C.

DNA preparation

Genomic DNA from each strain was isolated from overnight cultures using the DNeasy Blood and Tissue Kit (Qiagen, Valencia, CA), following the manufacturer's instructions. The extracted DNA was stored at -20˚C until used as a template for whole genome sequencing. The concentration was determined using a Qubit double-stranded DNA HS assay kit and a Qubit 2.0 fluorometer (Thermo Scientific), according to manufacturer's instructions.

MinION nanopore whole genome sequencing and contigs assembly

The genomes of the strains were sequenced using a MinION device (Nanopore, Oxford, UK), with FLO-MIN106 (R9.4) flow cells, according to the manufacturer's instructions, at > 50 X—average coverage. The sequencing libraries were prepared using either the 1D Genomic DNA by ligation kit (SQK-LSK108) library chemistry or the rapid sequencing kit (SQK-RAD002) according to the manufacturer's instructions. An exception was for the 1D ligation kit where we omitted the shearing step and the initial step was the End-prep step, since the DNA extraction step already sheared the DNA. The DNA input was 1 ug per DNA library for all 1D Genomic DNA by ligation kit (SQK-LSK108) and 0.4 ug per rapid sequencing kit (SQK-RAD002) library. Each prepared library was loaded into a FLO-MIN106 flowcell (R9.4) for a 48-hour run. *E. coli* strains CFSAN027346 and CFSAN027350 were sequenced in duplicate using the ligation kit, and a single sequencing of each strain was conducted using the rapid kit (Table 2). The base calling was performed using Albacore software (Nanopore, Oxford, UK). The fastq files were generated from the base called sequencing fast5 reads using Poretools [52]. Genomic sequence contigs were *de novo* assembled using default settings within the CANU program [53] v1.6. Determination of the overlapping regions of the chromosome and plasmids were carried out using Gepard [54]. The resultant assemblies from CANU were corrected for errors using Pilon [55] and the Nextera XT MiSeq data generated from those strains. Our de novo assembly and polishing pipeline are shown in Fig 1.

PacBio whole genome sequencing and contigs assembly

The strains were sequenced on the Pacific Biosciences (PacBio) *RS*II Sequencer, as previously described [2,18,48]. Specifically, we prepared the library using 10 μg genomic DNA that was sheared to a size of 20kb fragments by g-tubes (Covaris, Inc., Woburn, MA) according to the manufacturer's instruction. The SMRTbell 20-kb template library was constructed using DNA Template Prep Kit 1.0 with the 20-kb insert library protocol (Pacific Biosciences; Menlo Park, CA, USA). Size selection was performed with BluePippin (Sage Science, Beverly, MA). The library was sequenced using the P6/C4 chemistry on 3 single-molecule real-time (SMRT) cells with a 240-min collection protocol along with stage start.

Analysis of the sequence reads was implemented using SMRT Analysis 2.3.0. The best de novo assembly was established with the PacBio Hierarchical Genome Assembly Process (HGAP3.0) program using the continuous-long-reads from the four SMRT cells. The assemblies outputs from HGAP contains overlapping regions at the end which can be identified using dot plots in Gepard [54]. Genomes were checked manually for even sequencing coverage. Afterwards the improved consensus sequence was uploaded in SMRT Analysis 2.3.0. to determine the final consensus and accuracy scores using Quiver consensus algorithm [48].

MiSeq whole genome sequencing, contig assembly and annotation

The genomes of the strains were sequenced using an Illumina MiSeq sequencer (Illumina, San Diego, CA), with the 2x250 bp pair-end chemistry according to manufacturer's instructions, at

approximately 80X average coverage. The genome libraries were constructed using the Nextera XT DNA sample prep kit (Illumina). Genomic sequence contigs were *de novo* assembled using default settings within CLC Genomics Workbench v9.5.2 (QIAGEN) with a minimum contig size threshold of 500 bp in length.

Genome and plasmid annotations

We used RAST (http://rast.nmpdr.org/rast.cgi) for all annotations performed in this study [56].

In silico serotyping

The serotype of each strain analyzed in this study was confirmed using the genes deposited in the Center for Genomic Epidemiology (http://www.genomicepidemiology.org) for *E. coli* as part of their web-based serotyping tool (SerotypeFinder 1.1 - https://cge.cbs.dtu.dk/services/SerotypeFinder) [57,58].

In silico MLST phylogenetic analysis

The initial analysis and identification of the strains were performed using an *in silico E. coli* MLST approach, based on the information available at the *E. coli* MLST website Enterobase (http://enterobase.warwick.ac.uk/species/index/ecoli) and using Ridom SeqSphere+ software v2.4.0 (Ridom; Münster, Germany) (http://www.ridom.com/seqsphere). Seven housekeeping genes (*dnaE*, *gyrB*, *recA*, *dtdS*, *pntA*, *pyrC*, and *tnaA*), described previously for *E. coli* [15], were used for MLST analysis.

In silico determination of virulence genes

Virulence genes were determined as previously described [9] using the genes deposited in the Center for Genomic Epidemiology (http://www.genomicepidemiology.org) for *E. coli* as part of their VirulenceFinder 1.5 web-based tool (https://cge.cbs.dtu.dk/services/VirulenceFinder) [58], We used Ridom to batch screen our set of genomes for known virulence genes. S1 Table shows the 93 virulence genes analyzed by this method.

In silico identification of antimicrobial resistance genes

We identified the antimicrobial resistance (AMR) genes present in our sequenced genomes as previously described by Gonzalez-Escalona [9], using genes deposited in the Center for Genomic Epidemiology (http://www.genomicepidemiology.org) as part of their web-based Resfinder 2.1 tool (https://cge.cbs.dtu.dk/services/ResFinder) [59]. We used Ridom to batch screen our set of genomes for known AMR genes.

Comparisons of genomic synteny

The genomic synteny between PacBio and MinION data was determined using Mauve [39].

stx phage and T3SS identification and location

Prophages and prophage-like elements within the sequenced O26:H11 STEC strains were initially identified using the prophage-predicting PHASTER web server [38]. Next, we used the genomic island prediction web server IslandViewer3 to detect potential pathogenicity and genomic islands [60]. Each identified prophage, and prophage-like elementwas then examined,

using CLC Genomics Workbench 7.5, to locate nearby integrases and potential integration sites, which would confirm their status.

Phylogenetic relationship of the strains by cgMLST analysis

Due to intrinsic problems with the sequencing technology used by MinION sequencing–generation of many indels located mainly in areas of homopolymers tracks- we wanted to test the effectiveness of MinION genomes produced by our pipeline to estimate the phylogeny of those 3 STEC strains in a context of 155 other genomes of highly similar O26:H11 available at NCBI. All these O26 genomes from GenBank were generated by MiSeq or HighSeq (Illuimina), considered the "gold standard" for accurate genome sequence determination and SNP analyses. We used Ridom SeqSphere+ software v2.4.0 to assess the phylogenetic relationship of these strains, performing a core genome multilocus sequence typing (cgMLST) analysis as previously described for O26:H11 [9]. cgMLST uses the allele numbers of each loci to determine genetic distances and build the phylogenetic tree. We used O26:H11 strain 11368 (NC_013361.1) as the reference genome for generating the core genes for the phylogenetic tree and downloaded 195 genomes of *E. coli* O26:H11, available at NCBI, to build the phylogenetic tree. We used Nei's DNA distance method [61] for calculating the matrix of genetic distance, taking only the number of same/different alleles in the core genes into consideration. A Neighbor-Joining (NJ) tree [62] using the appropriate genetic distances was built after the cgMLST analysis for initial inspection of the placement of the new sequenced genomes by either technology. The SNPs were extracted from the core loci (1303) and the SNP matrix (5089 SNPs) was used to reconstruct the NJ phylogenetic tree. The use of SNPs allowed for determining a true phylogeny and to find informative SNPs in the dataset. The NJ SNP tree was rooted to strain 11128 (O111:H-).

Nucleotide sequence accession numbers

The SRA sequences of all three *E. coli* strains used in our study are available in GenBank under the accession numbers: MiSeq data (SRR8333591, SRR8333592, and SRR8333590), MinION data (SRR8335317, SRR8335318, and SRR8335317), and PacBio assemblies [CFSAN027343 (CP037943 and CP037944), CFSAN027346 (CP037945, CP037946, and CP037947), and CFSAN027350 (CP037941 and CP037942)].

Supporting information

S1 Fig. Annotation of plasmid 73 kb from Strain CFSAN027346 showing the antimicrobial resistance genes.
(DOCX)

S2 Fig. SNPs differences observed by a cgMLST analysis between the genomes generated by MiSeq, PacBio, and MinION. A) CFSAN027343, B) CFSAN027346.
(DOCX)

S3 Fig. Snapshot of the PHASTER [38] results for CFSAN027343 MinION chromosome.
(DOCX)

S4 Fig. Comparison of the synteny mapping of chromosomes generated by either MinION or PacBio, using the Sanger generated genome for STEC O26:H11 strain 11368 (AP010953) as reference with MAUVE. Each chromosome sequence is laid out in a horizontal track. Matching colors indicate homologous segments and are connected across genomes. Respective scales show the sequence coordinates in base pairs. A colored similarity plot is

shown for each genome, the height of which is proportional to the level of sequence identity in that region. Strains CFSAN027346 and CFSAN027350 are shown. Top sanger sequenced strain 11368, middle MinION sequenced strain, and bottom the same strain sequenced by PacBio.
(DOCX)

S5 Fig. Same phylogenetic tree as Fig 2A but on a larger scale to show the names of the strains.
(DOCX)

S1 Table. *E. coli* virulence genes tested by *in silico* virulence typing.
(DOCX)

S2 Table. Summary of ORFs in pCFSAN027350 as annotated by RAST [56].
(DOCX)

Acknowledgments

The authors thank Dr. Lili Fox Vélez for her scientific writing assistance on this manuscript. The views expressed in this article are those of the authors and do not necessarily reflect the official policy of the Department of Health and Human Services, the U.S. Food and Drug Administration (FDA), or the U.S. Government. Reference to any commercial materials, equipment, or process does not in any way constitute approval, endorsement, or recommendation by the FDA.

Author Contributions

Conceptualization: Narjol González-Escalona, Maria Hoffmann.

Data curation: Narjol González-Escalona.

Formal analysis: Narjol González-Escalona.

Funding acquisition: Narjol González-Escalona.

Investigation: Narjol González-Escalona.

Methodology: Narjol González-Escalona, Maria Hoffmann.

Project administration: Narjol González-Escalona.

Resources: Narjol González-Escalona.

Software: Narjol González-Escalona, Maria Hoffmann.

Supervision: Narjol González-Escalona.

Validation: Narjol González-Escalona.

Visualization: Narjol González-Escalona.

Writing – original draft: Narjol González-Escalona, Maria Hoffmann.

Writing – review & editing: Narjol González-Escalona, Marc A. Allard, Eric W. Brown, Shashi Sharma, Maria Hoffmann.

References

1. Allard MW, Bell R, Ferreira CM, Gonzalez-Escalona N, Hoffmann M, Muruvanda T et al. Genomics of foodborne pathogens for microbial food safety. Curr Opin Biotechnol. 2018; 49: 224–229. S0958-1669 (17)30139-8; https://doi.org/10.1016/j.copbio.2017.11.002 PMID: 29169072

2. Hoffmann M, Luo Y, Monday SR, Gonzalez-Escalona N, Ottesen AR, Muruvanda T et al. Tracing Origins of the *Salmonella* Bareilly Strain Causing a Food-borne Outbreak in the United States. J Infect Dis. 2016; 213: 502–508. jiv297 [pii]; https://doi.org/10.1093/infdis/jiv297 PMID: 25995194
3. Martinez-Urtaza J, van AR, Abanto M, Haendiges J, Myers RA, Trinanes J et al. Genomic Variation and Evolution of *Vibrio parahaemolyticus* ST36 over the Course of a Transcontinental Epidemic Expansion. MBio. 2017; 8. mBio.01425-17; https://doi.org/10.1128/mBio.01425-17 PMID: 29138301
4. Lorenz SC, Gonzalez-Escalona N, Kotewicz ML, Fischer M, Kase JA. Genome sequencing and comparative genomics of enterohemorrhagic *Escherichia coli* O145:H25 and O145:H28 reveal distinct evolutionary paths and marked variations in traits associated with virulence & colonization. BMC Microbiol. 2017; 17: 183. https://doi.org/10.1186/s12866-017-1094-3 PMID: 28830351
5. Strauss L, Stegger M, Akpaka PE, Alabi A, Breurec S, Coombs G et al. Origin, evolution, and global transmission of community-acquired *Staphylococcus aureus* ST8. Proc Natl Acad Sci U S A. 2017; 114: E10596–E10604. 1702472114; https://doi.org/10.1073/pnas.1702472114 PMID: 29158405
6. Deng X, Desai PT, den Bakker HC, Mikoleit M, Tolar B, Trees E et al. Genomic epidemiology of *Salmonella enterica* serotype Enteritidis based on population structure of prevalent lineages. Emerg Infect Dis. 2014; 20: 1481–1489. https://doi.org/10.3201/eid2009.131095 PMID: 25147968
7. Bletz S, Bielaszewska M, Leopold SR, Kock R, Witten A, Schuldes J et al. Evolution of enterohemorrhagic *Escherichia coli* O26 based on single-nucleotide polymorphisms. Genome Biol Evol. 2013; 5: 1807–1816. evt136 [pii]; https://doi.org/10.1093/gbe/evt136 PMID: 24105689
8. Gonzalez-Escalona N, Gavilan RG, Toro M, Zamudio ML, Martinez-Urtaza J. Outbreak of *Vibrio parahaemolyticus* Sequence Type 120, Peru, 2009. Emerg Infect Dis. 2016; 22: 1235–1237. https://doi.org/10.3201/eid2207.151896 PMID: 27315090
9. Gonzalez-Escalona N, Toro M, Rump LV, Cao G, Nagaraja TG, Meng J. Virulence Gene Profiles and Clonal Relationships of *Escherichia coli* O26:H11 Isolates from Feedlot Cattle as Determined by Whole-Genome Sequencing. Appl Environ Microbiol. 2016; 82: 3900–3912. AEM.00498-16 [pii]; https://doi.org/10.1128/AEM.00498-16 PMID: 27107118
10. Bielaszewska M, Mellmann A, Bletz S, Zhang W, Kock R, Kossow A et al. Enterohemorrhagic *Escherichia coli* O26:H11/H-: a new virulent clone emerges in Europe. Clin Infect Dis. 2013; 56: 1373–1381. cit055 [pii]; https://doi.org/10.1093/cid/cit055 PMID: 23378282
11. Bugarel M, Beutin L, Scheutz F, Loukiadis E, Fach P. Identification of genetic markers for differentiation of Shiga toxin-producing, enteropathogenic, and avirulent strains of *Escherichia coli* O26. Appl Environ Microbiol. 2011; 77: 2275–2281. AEM.02832-10 [pii]; https://doi.org/10.1128/AEM.02832-10 PMID: 21317253
12. Chase-Topping ME, Rosser T, Allison LJ, Courcier E, Evans J, McKendrick IJ et al. Pathogenic potential to humans of bovine *Escherichia coli* O26, Scotland. Emerg Infect Dis. 2012; 18: 439–448. https://doi.org/10.3201/eid1803.111236 PMID: 22377426
13. Dallman TJ, Byrne L, Launders N, Glen K, Grant KA, Jenkins C. 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. 2015; 143: 1672–1680. S0950268814002696 [pii]; https://doi.org/10.1017/S0950268814002696 PMID: 25316375
14. Fratamico PM, Yan X, Caprioli A, Esposito G, Needleman DS, Pepe T et al. The complete DNA sequence and analysis of the virulence plasmid and of five additional plasmids carried by Shiga toxin-producing *Escherichia coli* O26:H11 strain H30. Int J Med Microbiol. 2011; 301: 192–203. S1438-4221(10)00105-0 [pii]; https://doi.org/10.1016/j.ijmm.2010.09.002 PMID: 21212019
15. Wirth T, Falush D, Lan R, Colles F, Mensa P, Wieler LH et al. Sex and virulence in *Escherichia coli*: an evolutionary perspective. Mol Microbiol. 2006; 60: 1136–1151. MMI5172; https://doi.org/10.1111/j.1365-2958.2006.05172.x PMID: 16689791
16. Zweifel C, Cernela N, Stephan R. Detection of the emerging Shiga toxin-producing *Escherichia coli* O26:H11/H- sequence type 29 (ST29) clone in human patients and healthy cattle in Switzerland. Appl Environ Microbiol. 2013; 79: 5411–5413. AEM.01728-13 [pii]; https://doi.org/10.1128/AEM.01728-13 PMID: 23811503
17. Rhoads A, Au KF. PacBio Sequencing and Its Applications. Genomics Proteomics Bioinformatics. 2015; 13: 278–289. S1672-0229(15)00134-5; https://doi.org/10.1016/j.gpb.2015.08.002 PMID: 26542840
18. Hoffmann M, Muruvanda T, Allard MW, Korlach J, Roberts RJ, Timme R et al. Complete Genome Sequence of a Multidrug-Resistant *Salmonella enterica* Serovar Typhimurium var. 5- Strain Isolated from Chicken Breast. Genome Announc. 2013; 1. 1/6/e01068-13; https://doi.org/10.1128/genomeA.01068-13 PMID: 24356834

19. Loman NJ, Quick J, Simpson JT. A complete bacterial genome assembled de novo using only nanopore sequencing data. Nat Methods. 2015; 12: 733–735. nmeth.3444; https://doi.org/10.1038/nmeth.3444 PMID: 26076426

20. Giordano F, Aigrain L, Quail MA, Coupland P, Bonfield JK, Davies RM et al. De novo yeast genome assemblies from MinION, PacBio and MiSeq platforms. Sci Rep. 2017; 7: 3935. https://doi.org/10.1038/s41598-017-03996-z PMID: 28638050

21. Quick J, Loman NJ, Duraffour S, Simpson JT, Severi E, Cowley L et al. Real-time, portable genome sequencing for Ebola surveillance. Nature. 2016; 530: 228–232. https://doi.org/10.1038/nature16996 PMID: 26840485

22. Hoenen T, Groseth A, Rosenke K, Fischer RJ, Hoenen A, Judson SD et al. Nanopore Sequencing as a Rapidly Deployable Ebola Outbreak Tool. Emerg Infect Dis. 2016; 22: 331–334. https://doi.org/10.3201/eid2202.151796 PMID: 26812583

23. Greninger AL, Naccache SN, Federman S, Yu G, Mbala P, Bres V et al. Rapid metagenomic identification of viral pathogens in clinical samples by real-time nanopore sequencing analysis. Genome Medicine. 2015; 7: 99. https://doi.org/10.1186/s13073-015-0220-9 PMID: 26416663

24. Benitez-Paez A, Portune KJ, Sanz Y. Species-level resolution of *16S rRNA* gene amplicons sequenced through the MinION portable nanopore sequencer. Gigascience. 2016; 5: 4. https://doi.org/10.1186/s13742-016-0111-z PMID: 26823973

25. Brown BL, Watson M, Minot SS, Rivera MC, Franklin RB. MinION nanopore sequencing of environmental metagenomes: a synthetic approach. Gigascience. 2017. 3051932; https://doi.org/10.1093/gigascience/gix007 PMID: 28327976

26. Judge K, Harris SR, Reuter S, Parkhill J, Peacock SJ. Early insights into the potential of the Oxford Nanopore MinION for the detection of antimicrobial resistance genes. J Antimicrob Chemother. 2015; 70: 2775–2778. dkv206; https://doi.org/10.1093/jac/dkv206 PMID: 26221019

27. van der Helm E, Imamovic L, Hashim Ellabaan MM, van SW, Koza A, Sommer MOA. Rapid resistome mapping using nanopore sequencing. Nucleic Acids Res. 2017; 45: e61. gkw1328; https://doi.org/10.1093/nar/gkw1328 PMID: 28062856

28. Tilgner H, Jahanbani F, Blauwkamp T, Moshrefi A, Jaeger E, Chen F et al. Comprehensive transcriptome analysis using synthetic long-read sequencing reveals molecular co-association of distant splicing events. Nat Biotechnol. 2015; 33: 736–742. nbt.3242; https://doi.org/10.1038/nbt.3242 PMID: 25985263

29. Simpson JT, Workman RE, Zuzarte PC, David M, Dursi LJ, Timp W. Detecting DNA cytosine methylation using nanopore sequencing. Nat Methods. 2017; 14: 407–410. nmeth.4184; https://doi.org/10.1038/nmeth.4184 PMID: 28218898

30. Loman NJ, Watson M. Successful test launch for nanopore sequencing. Nat Methods. 2015; 12: 303–304. nmeth.3327; https://doi.org/10.1038/nmeth.3327 PMID: 25825834

31. Lu H, Giordano F, Ning Z. Oxford Nanopore MinION Sequencing and Genome Assembly. Genomics Proteomics Bioinformatics. 2016; 14: 265–279. S1672-0229(16)30130-9; https://doi.org/10.1016/j.gpb.2016.05.004 PMID: 27646134

32. Jain M, Olsen HE, Paten B, Akeson M. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. Genome Biology. 2016; 17: 239. https://doi.org/10.1186/s13059-016-1103-0 PMID: 27887629

33. Norris AL, Workman RE, Fan Y, Eshleman JR, Timp W. Nanopore sequencing detects structural variants in cancer. Cancer Biol Ther. 2016; 17: 246–253. https://doi.org/10.1080/15384047.2016.1139236 PMID: 26787508

34. Loose M, Malla S, Stout M. Real-time selective sequencing using nanopore technology. Nat Methods. 2016; 13: 751–754. nmeth.3930; https://doi.org/10.1038/nmeth.3930 PMID: 27454285

35. Jenkins C, Evans J, Chart H, Willshaw GA, Frankel G. *Escherichia coli* serogroup O26—a new look at an old adversary. J Appl Microbiol. 2008; 104: 14–25. JAM3465 [pii]; https://doi.org/10.1111/j.1365-2672.2007.03465.x PMID: 18171379

36. Nataro JP, Kaper JB. Diarrheagenic *Escherichia coli*. Clin Microbiol Rev. 1998; 11: 142–201. PMID: 9457432

37. Gonzalez-Escalona N, Toro M, Rump LV, Cao G, Nagaraja TG, Meng J. Virulence Gene Profiles and Clonal Relationships of *Escherichia coli* O26:H11 Isolates from Feedlot Cattle as Determined by Whole-Genome Sequencing. Appl Environ Microbiol. 2016; 82: 3900–3912. AEM.00498-16 [pii]; https://doi.org/10.1128/AEM.00498-16 PMID: 27107118

38. Arndt D, Grant JR, Marcu A, Sajed T, Pon A, Liang Y et al. PHASTER: a better, faster version of the PHAST phage search tool. Nucleic Acids Res. 2016; 44: W16–W21. gkw387; https://doi.org/10.1093/nar/gkw387 PMID: 27141966

39. Darling ACE, Mau B, Blattner FR, Perna NT. Mauve: Multiple alignment of conserved genomic sequence with rearrangements. Genome Res. 2004; 14: 1394–1403. https://doi.org/10.1101/gr.2289704 PMID: 15231754

40. Ogura Y, Gotoh Y, Itoh T, Sato MP, Seto K, Yoshino S et al. Population structure of *Escherichia coli* O26: H11 with recent and repeated *stx2* acquisition in multiple lineages. Microb Genom. 2017; 3. https://doi.org/10.1099/mgen.0.000141 PMID: 29208163

41. Ishijima N, Lee KI, Kuwahara T, Nakayama-Imaohji H, Yoneda S, Iguchi A et al. Identification of a New Virulent Clade in Enterohemorrhagic *Escherichia coli* O26:H11/H- Sequence Type 29. Sci Rep. 2017; 7: 43136. srep43136; https://doi.org/10.1038/srep43136 PMID: 28230102

42. Garmendia J, Ren Z, Tennant S, Midolli Viera MA, Chong Y, Whale A et al. Distribution of tccP in clinical enterohemorrhagic and enteropathogenic *Escherichia coli* isolates. J Clin Microbiol. 2005; 43: 5715–5720. 43/11/5715; https://doi.org/10.1128/JCM.43.11.5715-5720.2005 PMID: 16272509

43. Garmendia J, Frankel G, Crepin VF. Enteropathogenic and enterohemorrhagic *Escherichia coli* infections: translocation, translocation, translocation. Infect Immun. 2005; 73: 2573–2585. 73/5/2573; https://doi.org/10.1128/IAI.73.5.2573-2585.2005 PMID: 15845459

44. Garmendia J, Phillips AD, Carlier MF, Chong Y, Schuller S, Marches O et al. TccP is an enterohaemorrhagic *Escherichia coli* O157:H7 type III effector protein that couples Tir to the actin-cytoskeleton. Cell Microbiol. 2004; 6: 1167–1183. CMI459; https://doi.org/10.1111/j.1462-5822.2004.00459.x PMID: 15527496

45. Cepeda-Molero M, Berger CN, Walsham ADS, Ellis SJ, Wemyss-Holden S, Schuller S et al. Attaching and effacing (A/E) lesion formation by enteropathogenic *E. coli* on human intestinal mucosa is dependent on non-LEE effectors. PLoS Pathog. 2017; 13: e1006706. https://doi.org/10.1371/journal.ppat.1006706 PPATHOGENS-D-17-00629. PMID: 29084270

46. Etcheverria AI, Padola NL. Shiga toxin-producing *Escherichia coli*: factors involved in virulence and cattle colonization. Virulence. 2013; 4: 366–372. 24642; https://doi.org/10.4161/viru.24642 PMID: 23624795

47. Lorenz SC, Kotewicz ML, Hoffmann M, Gonzalez-Escalona N, Fischer M, Kase JA. Complete Genome Sequences of Four Enterohemolysin-Positive (*ehxA*) Enterocyte Effacement-Negative Shiga Toxin-Producing *Escherichia coli* Strains. Genome Announc. 2016; 4. 4/5/e00846-16 [pii]; https://doi.org/10.1128/genomeA.00846-16 PMID: 27587806

48. Chin CS, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C et al. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat Methods. 2013; 10: 563–569. nmeth.2474; https://doi.org/10.1038/nmeth.2474 PMID: 23644548

49. Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. Nat Biotechnol. 2018; 36: 338–345. nbt.4060; https://doi.org/10.1038/nbt.4060 PMID: 29431738

50. Bonanno L, Loukiadis E, Mariani-Kurkdjian P, Oswald E, Garnier L, Michel Vr et al. Diversity of Shiga Toxin-Producing *Escherichia coli* (STEC) O26:H11 Strains Examined via *stx* Subtypes and Insertion Sites of Stx and EspK Bacteriophages. Appl Environ Microbiol. 2015; 81: 3712–3721. https://doi.org/10.1128/AEM.00077-15 PMID: 25819955

51. Delannoy S, Mariani-Kurkdjian P, Webb HE, Bonacorsi S, Fach P. The Mobilome; A Major Contributor to *Escherichia coli* stx2-Positive O26:H11 Strains Intra-Serotype Diversity. Front Microbiol. 2017; 8: 1625. https://doi.org/10.3389/fmicb.2017.01625 PMID: 28932209

52. Loman NJ, Quinlan AR. Poretools: a toolkit for analyzing nanopore sequence data. Bioinformatics. 2014; 30: 3399–3401. btu555; https://doi.org/10.1093/bioinformatics/btu555 PMID: 25143291

53. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. 2017; 27: 722–736. gr.215087.116; https://doi.org/10.1101/gr.215087.116 PMID: 28298431

54. Krumsiek J, Arnold R, Rattei T. Gepard: a rapid and sensitive tool for creating dotplots on genome scale. Bioinformatics. 2007; 23: 1026–1028. btm039; https://doi.org/10.1093/bioinformatics/btm039 PMID: 17309896

55. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One. 2014; 9: e112963. https://doi.org/10.1371/journal.pone.0112963 PONE-D-14-38252. PMID: 25409509

56. Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA et al. The RAST Server: rapid annotations using subsystems technology. BMC Genomics. 2008; 9: 75. 1471-2164-9-75; https://doi.org/10.1186/1471-2164-9-75 PMID: 18261238

57. Joensen KG, Tetzschner AM, Iguchi A, Aarestrup FM, Scheutz F. Rapid and Easy In Silico Serotyping of *Escherichia coli* Isolates by Use of Whole-Genome Sequencing Data. J Clin Microbiol. 2015; 53: 2410–2426. JCM.00008-15 [pii]; https://doi.org/10.1128/JCM.00008-15 PMID: 25972421

58. Joensen KG, Scheutz F, Lund O, Hasman H, Kaas RS, Nielsen EM et al. Real-time whole-genome sequencing for routine typing, surveillance, and outbreak detection of verotoxigenic *Escherichia coli*. J Clin Microbiol. 2014; 52: 1501–1510. JCM.03617-13 [pii]; https://doi.org/10.1128/JCM.03617-13 PMID: 24574290

59. Zankari E, Hasman H, Cosentino S, Vestergaard M, Rasmussen S, Lund O et al. Identification of acquired antimicrobial resistance genes. J Antimicrob Chemother. 2012; 67: 2640–2644. dks261 [pii]; https://doi.org/10.1093/jac/dks261 PMID: 22782487

60. Dhillon BK, Laird MR, Shay JA, Winsor GL, Lo R, Nizam F et al. IslandViewer 3: more flexible, interactive genomic island discovery, visualization and analysis. Nucleic Acids Res. 2015; 43: W104–W108. gkv401; https://doi.org/10.1093/nar/gkv401 PMID: 25916842

61. Nei M, Tajima F, Tateno Y. Accuracy of estimated phylogenetic trees from molecular data. II. Gene frequency data. J Mol Evol. 1983; 19: 153–170. PMID: 6571220

62. Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol. 1987; 4: 406–425. https://doi.org/10.1093/oxfordjournals.molbev.a040454 PMID: 3447015

63. Stothard P, Wishart DS. Circular genome visualization and exploration using CGView. Bioinformatics. 2005; 21: 537–539. bti054 [pii]; https://doi.org/10.1093/bioinformatics/bti054 PMID: 15479716