GigaScience

Comparison of single nucleotide variants identified by Illumina and Oxford Nanopore technologies in the context of a potential outbreak of Shiga Toxin Producing Escherichia coli

--Manuscript Draft--

Manuscript Number:	GIGA-D-19-00070R1
Full Title:	Comparison of single nucleotide variants identified by Illumina and Oxford Nanopore technologies in the context of a potential outbreak of Shiga Toxin Producing Escherichia coli
Article Type:	Research
Funding Information:	Oxford Nanopore Technologies — Not applicable
Abstract:	Background We aimed to compare Illumina and Oxford Nanopore Technology (ONT) sequencing data from the two isolates of STEC O157:H7 to determine whether concordant single nucleotide variants were identified and whether inference of relatedness was consistent with the two technologies. Results For the Illumina workflow, the time from DNA extraction to availability of results, was approximately 40 hours in comparison to the ONT workflow where serotyping, Shiga toxin subtyping variant identification were available within seven hours. After optimisation of the ONT variant filtering, on average 95% of the discrepant positions between the technologies were accounted for by methylated positions found in the described 5-Methylcytosine motif sequences, CC(A/T)GG. Of the few discrepant variants (6 and 7 difference for the two isolates) identified by the two technologies, it is likely that both methodologies contain false calls. Conclusions Despite these discrepancies, Illumina and ONT sequences from the same case were placed on the same phylogenetic location against a dense reference database of STEC O157:H7 genomes sequenced using the Illumina workflow. Robust SNP typing using MinION-based variant calling is possible and we provide evidence that the two technologies can be used interchangeably to type STEC O157:H7 in a public health setting.
Corresponding Author:	Timothy Dallman UNITED KINGDOM
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	
Corresponding Author's Secondary Institution:	
First Author:	David R Greig
First Author Secondary Information:	
Order of Authors:	David R Greig Claire Jenkins Saheer Gharbia Timothy J Dallman
Order of Authors Secondary Information:	
Response to Reviewers:	Reviewer reports: Reviewer #1: This manuscript describes a comparative analysis of Illumina and Nanopore sequencing, evaluating their usefulness for phylogenetic analysis, and for

identifying genetic variants in outbreak situations.

The outcome of the research is somewhat surprising, given the expectation that Illumina sequencing represents the current gold-standard in sequencing accuracy. When eliminating systematic variation in base sequences caused by methylation, Nanopore sequencing appears to have similar accuracy to Illumina sequencing for the purpose of variant categorisation. When methylation is considered as a important feature, Nanopore sequencing demonstrates both a greater detection ability, and a faster turnaround time compared to Illumina sequencing. I am pleased to note that the supporting data was available at the time I carried out my review, and also pleased to be given the opportunity to approve this manuscript for publication.

I was specifically asked by the editors to state whether this represents "the state-of-the-art in terms of what this platform can do." It would be underselling the impact of these results to say no, that the current basecalling technology is better than what is presented in this paper. Due to the rapid advancement of nanopore sequencing technology in hardware, software, and chemistry, the yield and quality of results obtained from nanopore sequencing will be better than what is in *any* publication, even at the time when a
manuscript is submitted for review.

I recall seeing (and commenting) on David, Claire, Kathie, and Timothy's poster presented in April 2018, which seems to have been a similar (if not the same) study [https://twitter.com/gingerdavid92/status/987947325086666753]. This was the first study I'd seen that explicitly compared Illumina and Nanopore sequencing for phylogenetics [I accept there may be others
that I haven't seen], and I'm pleased to see that they have incorporated an explicit analysis of methylation signals since then.

People have previously looked at phylogenetic trees for outbreak tracking with Nanopore sequencing
(e.g. https://doi.org/10.2807%2F1560-7917.ES.2018.23.12.17-00140 [essentially cited in ref#10]), at accuracy estimates for Nanopore basecalling (e.g. https://doi.org/10.1101/543439), at hybrid isolate assembly from
barcoded Nanopore and Illumina reads (e.g. https://doi.org/10.1099%2Fmgen.0.000132 [cited]), and at comparing clinical turnaround time for Nanopore vs Illumina (e.g. https://doi.org/10.1128/JCM.02483-16), but this paper puts it all together into something that is still of substantial interest to the research community, as demonstrated by the social media impact of their preprint (https://doi.org/10.1101/570192).

In short, this manuscript is an excellent demonstration of what nanopore sequencing is capable of, represents the state-of-the-art (as I understand it) for public health investigations as presented in published papers, and I look forward to seeing more studies like this in the future.

Additional comments / questions:

1.Results, Tables 1/2 line 194-200 - Could you please either add in the legend that these SNPs were homoplasmic (very unlikely for ONT, somewhat possible for Illumina), or add the depth of the reference SNP bases to the table?

We have added the depth of sequencing for the final discrepant SNPs in to Table 2. We have identified what SNPs in the were homoplasmic (5/7 Illumina variants) and a line in the text.

2.Methods, line 348 - These were barcoded reads that were processed through Porechop, which I understand can identify and filter out chimeric reads. Do you know how many reads were chimeric (we've typically observed <0.5% chimeric reads from rapid adapter preps, about 4% from ligation preps)?

We have added a line in the methods with these figures.

3.Discussion, line 239 - It is interesting to see from Figure 1 that all the nanopore data analysis was completed before the sequencing run had ended. Maybe this could be emphasised here: "within 377 minutes (i.e. over 20 hours *before* the sequencing run was scheduled to finish)."

We have added a short statement emphasising the difference between technologies.

4.Discussion, lines 250-259 - The final sentence doesn't seem to match the general idea of this paragraph. The paragraph is about single-base accuracy for single molecules (note: Illumina never sequences a single molecule to generate a base call), whereas the last sentence is about phylogenetics. I'd be happier if this paragraph were deleted entirely, as phylogenetics and error are also discussed in the next paragraph.

We have removed the last sentence from this paragraph. We think it is important to keep the current limitations to show what was state of the art at the time of this publication

5.Discusion, line 283 - "long reads... workflow is" -> "longreads... workflow are"

This has now been corrected.

6.Discussion, line 303 - "up-dates" -> "updates"

This has now been corrected.

7.Figure 1 - Why were different methods used for DNA extraction (i.e. Promega Wizard vs Manual lysis / Qiagen Qiasymphony)?

The Qiasymphony utilises a magnetic beads in beating protocol that causes DNA fragmentation leading to a decrease in high molecular weight DNA molecules and thereby sub-optimal for long read sequencing. Therefore, a commercial gDNA extraction kit with modifications to attempt to keep the DNA integrity as high as possible and thus generate longer reads. We have added a sentence in the text to reflect the motivation of DNA extraction method.

8.Figure 2 - The numbers are difficult to read. Could the axis text be made larger?

We have enlarged the text for the axis for both graphs in Figure 2.

9.Figure 4 - This should be a line graph (similar to figure 3). The points represent sampling of potential cutoff scores along a continuous distribution, and the score represents a single value rather than count data.

We have modified the figure accordingly

10.Figure 5 - Table 1 (Line 165-166) mentions that the total number of discrepant variants for case A and B is 266 and 101 respectively. This doesn't match the percentages and totals represented in Figure 5. I would expect that the Total line for A/B in Figure 5 should be 97.7% and 93% respectively, indicating that transitions comprised that proportion of the total variants. It would be useful to refer back to Tables 1 & 2 in the text for the other discrepant variants.

We have clarified this in the legend in figure 5. In table 1 we are showing the total number of discrepant positions within both Illumina and ONT vcfs, however in figure 5 we are discussing only the variant positions in the ONT data alone that were determined to be methylated. Figure 5 should equate to the row titled '# of discrepant variants with methylated positions masked' in Table 1.

11.Figure 6 - What do the numbers represent? It is not clear from the figure legend. These are presumably not bootstrap values, as they have a consistent ordering from top to bottom.

These values represent each respective case's SNP differences between the Illumina

and ONT as demonstrated on the tree. We have updated the figure (6) legend to now include this clarification.

General questions:

Given that the Nanopore technology has improved in a number of different areas since this investigation was carried out (e.g. 9.4.1 Series D Flow cells, Field sequencing kit and/or RBK004, flip-flop basecaller), what (if anything) would be done differently if you had the opportunity to do this again?

Of your suggestions, the only one likely to make a significant difference would be re-basecalling with the most up to date flip-flop basecaller, we would hypothesise that this would reduce the number of SNP differences if it accounts for methylation better than previous basecalling algorithms. Another advancement is the new R10 pores which aims to improve the consensus accuracy is about 99.999% which again would reduce the number of total SNP differences but will most likely not account for methylation (unless a trained basecaller is also developed with these pores to account for methylation).

Are the assemblies available? I can't see anything about the assemblies in the "Availability" section.

We have now uploaded our assemblies to NCBI and have added a comment (with accession numbers) in the “Availability of supporting data” section (lines 409-416).

Reviewer #2: In their paper, Greig et al. compare the performance of Oxford Nanopore Technology (ONT) and Illumina sequencing in identifying, subtyping and classifying clinical isolates in the context of outbreak investigation. This study is of considerable interest to both research and medical communities in two key aspects. Firstly, it provides an in-depth assessment of the performance of ONT versus the current sequencing standard Illumina Technology, and identifies the main mechanistic reason for discrepancies between the technologies (DNA methylation), which would be of value in optimizing and improving Nanopore analysis workflows. Secondly, they demonstrate that their real-time ONT analysis pipeline is able to rapidly provide diagnostic calls (species identification, serotyping etc) with comparable accuracy to Illumina sequencing in a fraction of the time, which has major potential applications in outbreak investigation.

Given the utility of this study, I would be happy to recommend it for publication - please consider the following points to possibly improve it further.

1) Are the methods appropriate to the aims of the study, are they well described, and are necessary controls included?
1.The sample size of 2 isolates is small, but justified given the context of the study (2 urgent cases of children with HUS admitted to the same hospital on the same night). However, if the authors have any other isolates sequenced (in particular, a reference isolate closely related to the reference genome) that allow for the comparison of Illumina/ONT, they could include it as supplemental information to improve the robustness of their assessment.

Currently we have only processed these two STEC O157:H7 samples using this methodology, we are hoping to follow up this manuscript in the future on a larger set of samples.

2.Run parameters for bioinformatics tools are well optimized and described. It would also be of major help to the community if the authors are willing to share the code for their real-time analysis pipeline.

Each component/tool was run individually during this study. We have not developed an automated pipeline though this is something we plan to develop in the future.

3.The authors adequately discuss the limitations of ONT relative to Illumina sequencing

with respect to their application in rapid diagnosis.
Fig 1/Methods - In the comparison of the ONT/Illumina workflows, we note that two different DNA extraction methods are used (manual + QiaSymphony cleanup for Illumina, Promega Wizard Genomic DNA Purification for ONT). Are the methods interchangeable for the purposes of the workflow?

See point 7 for reviewer 1

2) Are the conclusions adequately supported by the data shown?
Analyses are generally robust and well-supported, but we would like clarification on the following points:

4.Line 214 - When comparing the case B ONT sequence with the 3 concurrent outbreak isolates, was it compared against Illumina sequences, or Nanopore sequences? If the comparison was between ONT and Illumina sequences, the discordance might arise from differences in the base-calling/software methods, and might disappear if all isolates were sequenced with ONT and compared directly (or would the high error rate preclude a valid comparison?) Please clarify and comment.

The outbreak case B sequenced via ONT was compared to Illumina sequenced isolates (and the equivalent case B sample sequenced via Illumina). We believe that the observed differences are inherent errors in both technologies. Our hypothesis is in agreement with yours, that comparing ONT to ONT sequenced outbreak strains would remove these discrepancies.

5.Line 216-218 - Given that 7 SNPs is not too dramatic a difference one could still make the case that the cases are quite plausibly linked. Would you be able to set an approximate SNV threshold for concluding genetic linkage?

Currently we use 5 SNPs as a proxy to infer genetic linkage or sharing the same epidemiological source with Illumina sequenced strains, through extensive validation from sequencing known outbreaks. With ONT sequencing and due to the lack of background sequenced samples we are unable to comment on an “appropriate” SNP threshold. We would have to take into account comparing ONT to Illumina data, Illumina to Illumina data and ONT to ONT data to decide if each type of comparison requires a different threshold or if we could set a general one to cover all comparisons.

3) Please indicate the quality of language in the manuscript. Does it require a heavy editing for language and clarity?

6.The language of the manuscript is quite clear. Please correct "manufactures instructions" to "manufacturer's instructions".

This has now been corrected at each use.

7.I also feel that the title of the manuscript downplays the speed and relative accuracy of the ONT diagnostic pipeline - the focus of the title should not be on the comparison of SNVs, but rather the comparison of the overall performance of the two methods. A title reflecting this and highlighting the rapid, real-time analysis capability of ONT-based diagnostics would be able to better capture reader interest and increase the impact of the manuscript.

8.Similarly, the abstract should be edited to emphasize the speed and real-time analysis capability.

We feel that the current title is appropriate for this study as the emphasis was that in conjunction with the nanopore being a rapid, real-time portable sequencer the current dogma is that variant calling is currently out of scope for this technology due to the high error rate. This manuscript refutes that dogma.

4) Are you able to assess all statistics in the manuscript, including the appropriateness of statistical tests used?
Yes - our group has experience analysing similar datasets. The precision/recall analysis performed is straightforward and appropriate.

(This manuscript was co-reviewed with Weizhen Xu, a postdoctoral fellow in my research group)

Reviewer #3: The authors examine the feasibility of using Nanopore sequencing to characterise single nucleotide variants in clinically relevant outbreaks. The authors present an interesting and relevant comparison of sequencing technologies given the rapid uptake of WGS as a clinical diagnostic tool. The data set is clearly described and accession code for all data submitted to the SRA are available in the manuscript and online. The methods are well described and all software/Bioinformatic tools are available online.

1.Although it can be addressed, my main concern with the manuscript is that the research was part funded by Oxford Nanopore. I believe the authors have not fully addressed the limitations of the Nanopore sequencing technology e.g. Cost, variability in throughput, etc. I? have highlighted some of these issues below. Platform limitations will also need to be included in the discussion

We have been more explicit to the funding received by Oxford Nanopore in the Acknowledgments. We have a paragraph on the limitations of ONT sequencing in terms of read accuracy and therefore variant detection which is the focus of this paper.

2.In the abstract and results the authors make a comparison of Illumina and ONT workflows of the time taken from DNA extraction to availability of results. The authors state that typing data (Shiga toxin subtyping and serotyping) was available within 7 hours while with Illumina it took ~40 hours to get these results. How do these time frames compare to standard laboratory based typing techniques that might be available in a diagnostic/pathology lab?

This paper covers the comparison between current methods deployed in the national reference laboratory – WGS by Illumina – with an alternative sequencing methodology ONT. It is out of scope of the paper to consider methods deployed in diagnostics e.g PCR

3.I would like to see a comparison of the sequencing costs for Illumina and Nanopore sequencing. A comparison against standard laboratory based typing techniques might also be beneficial to a broader audience (i leave that to the authors to decide).

Costing the sequencing technologies is not the focus of this paper and a would need to be a paper in its' own right considering labour / non-labour cost, deployment models etc. Also the value of such a comparison is incredibly time limited.

4.The authors state that the genetic relatedness of isolates could be determined at ~6 hours. I assume by genetic relatedness we are talking about variant/SNP typing. Can the authors explain how variants could be determined at 6 hours yet a MLST profile for Case B could not be determined until 10 hours had passed? Surely an inability to determine a MLST type indicates that the genome has not yet been sufficiently covered and it therefore unsuitable for variant typing.

Our SNP typing process requires an average genome coverage of 30x, the ONT sequencing took 6 hours to achieve this. As a result, as soon as 30x coverage is passed, this process can begin. Whereas, when sequencing the seven MLST genes, we are looking for enough coverage of those genes so that krocus can give us a confirmed result. This typically takes much longer, the last read was aligned to the seventh MLST gene at about 10 hours to then generate a full MLST result.

5.Additionally, in order to be cost effective multiple samples would need to sequenced on the same flowcell at the same time. Can the authors comment on what impacted multiplexing might have on the time frames described here?

This is correct, it is standard to multiplex several isolates per flow-cell. The higher the degree of multiplexing the increased pore competition we would expect the time to

	receive results per sample to increase. We have not performed this comparison. We have added this to the discussion 6.Can the authors comment as to why a number of regions (all describes as prophage with the exception of 1 in case A) were only present in the ONT-only chromosome assemblies? I find it odd that these regions were not present in the Illumina sequence data and are also absent from the hybrid assembly. can the authors comment on why this might be the case and what impact that might have for genome sequencing and assembly strategies moving forward? The main reason for the smaller assemblies in the Illumina and hybrid approaches is the large amount of paralogous sequences in STEC O157 encoded on cryptic phage. These sequences (which are longer then the Illumina read length) collapse into a single contig or are broken up into many small contigs with only a single copy when in reality they are multi-copy in the genome. This results in smaller genomes when Illumina reads alone or as a hybrid. 7.With regards to the genome assemblies of case A and Case B can the authors provide information on the number of erroneous indels that were present in the Nanopore assemblies? I assume these errors were polished out but did the authors only use Nanopore sequence data or was Illumina data also required.? To keep the comparison as true as possible we only polished the ONT assembly with ONT data using Nanopolish. It is difficult in this case to quantify how many of the indels associated in the ONT assembly are correct or conversely incorrect in the Illumina assembly as many fall in prophage regions. 8.Line 129: Remove the MLST allele numbers This has now been corrected. 9.Line 385: form -> for(?) This has now been corrected. 10.Table 1 is very hard to interrupt. Consider restructuring the table. We have reformatted the table to make the breakdown of SNPs clearer.
Additional Information:	
Question	**Response**
Are you submitting this manuscript to a special series or article collection?	No
Experimental design and statistics Full details of the experimental design and statistical methods used should be given in the Methods section, as detailed in our Minimum Standards Reporting Checklist. Information essential to interpreting the data presented should be made available in the figure legends. Have you included all the information requested in your manuscript?	Yes

Resources A description of all resources used, including antibodies, cell lines, animals and software tools, with enough information to allow them to be uniquely identified, should be included in the Methods section. Authors are strongly encouraged to cite Research Resource Identifiers (RRIDs) for antibodies, model organisms and tools, where possible. Have you included the information requested as detailed in our Minimum Standards Reporting Checklist?	Yes
Availability of data and materials All datasets and code on which the conclusions of the paper rely must be either included in your submission or deposited in publicly available repositories (where available and ethically appropriate), referencing such data using a unique identifier in the references and in the “Availability of Data and Materials” section of your manuscript. Have you have met the above requirement as detailed in our Minimum Standards Reporting Checklist?	Yes

Comparison of single nucleotide variants identified by Illumina and Oxford Nanopore technologies in the context of a potential outbreak of Shiga Toxin Producing *Escherichia coli*.

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

National Infection Service, Public Health England, London, NW9 5EQ.

*Corresponding author.

Author details:

David R Greig:

Email – David.Greig@phe.gov.uk

ORCID: 0000-0001-9436-067X

Claire Jenkins:

Email - Claire.Jenkins1@phe.gov.uk

ORCID: 0000-0001-8600-9169

Saheer Gharbia

Email – Saheer.Gharbia@phe.gov.uk

Timothy J Dallman:

Email - Tim.Dallman@phe.gov.uk

ORCID: 0000-0001-7105-2543

Keywords –Oxford Nanopore, Illumina, Variant calling, STEC, outbreak.

Abstract

Background

We aimed to compare Illumina and Oxford Nanopore Technology (ONT) sequencing data from the two isolates of STEC O157:H7 to determine whether concordant single nucleotide variants were identified and whether inference of relatedness was consistent with the two technologies.

Results

For the Illumina workflow, the time from DNA extraction to availability of results, was approximately 40 hours in comparison to the ONT workflow where serotyping, Shiga toxin subtyping variant identification were available within seven hours. After optimisation of the ONT variant filtering, on average 95% of the discrepant positions between the technologies were accounted for by methylated positions found in the described 5-Methylcytosine motif sequences, CC(A/T)GG. Of the few discrepant variants (6 and 7 difference for the two isolates) identified by the two technologies, it is likely that both methodologies contain false calls.

Conclusions

Despite these discrepancies, Illumina and ONT sequences from the same case were placed on the same phylogenetic location against a dense reference database of STEC O157:H7 genomes sequenced using the Illumina workflow. Robust SNP typing using MinION-based variant calling is possible and we provide evidence that the two technologies can be used interchangeably to type STEC O157:H7 in a public health setting.

Background

Shiga toxin producing *Escherichia coli* (STEC) O157:H7 is a zoonotic, foodborne pathogen defined by the presence of phage-encoded Shiga toxin genes (*stx*) [1]. Disease symptoms range from mild through to severe bloody diarrhoea, often accompanied by fever, abdominal cramps and vomiting [2]. The infection can progress to Haemolytic Uremic Syndrome (HUS), characterized by kidney failure and/or cardiac and neurological complications [3,4]. Transmission from an animal reservoir, mainly ruminants, occurs by direct contact with animals or their environment, or by the consumption of contaminated food products with reported vehicles including beef and lamb meat, dairy products, raw vegetables and salad [2,4].

STEC O157:H7 belongs to multi-locus sequence type clonal complex (CC) 11, with all but a small number of variants belonging to sequence type ST11. CC11 comprises three main lineages (I, II and I/II) and seven sub-lineages (Ia, Ib, Ic, IIa, IIb, IIc and I/II) [5]. There are two types of Shiga toxin, Stx1 and Stx2. Stx1 has four subtypes (1a-1d) and Stx2 has seven subtypes (2a-2g). Subtypes 1a, 2a, 2c, and rarely 2d, are found in STEC O157:H7. Strains harbouring *stx2a* are significantly associated with cases that develop HUS [2,6]. As well as harbouring *stx* encoding prophage, STEC O157:H7 has an additional prophage repertoire accounting for at least 20% of the chromosome.

The implementation of whole genome sequencing (WGS) data for typing STEC has improved the detection and management of outbreaks of foodborne disease [6]. Single nucleotide polymorphism (SNP) typing offers an unprecedented level of strain discrimination and can be used to quantify the genetic relatedness between groups of genomes. In general, for clonal bacteria, the fewer polymorphisms identified between pairs of strains, the less time since divergence from a common ancestor and therefore the increased likelihood that they are from the same source population. Therefore, it is paramount that variant detection for typing is accurate, highly specific and concentrated on positions of neutral evolution to ensure the correct interpretation of the sequence data within the epidemiological context of an outbreak. It has been previously shown that different bioinformatics analysis approaches for variant identification exhibit detection variability [7,8]. It is therefore important that within a particular analysis, workflow parameters to filter identified variants to achieve optimum sensitivity and specificity are appropriately optimised.

Short read sequencing platforms, such as those provided by Illumina, have been adopted by public health agencies for infectious disease surveillance worldwide [9] and have proved to be a robust and accurate method for quantifying relatedness between bacterial genomes. High-throughput Illumina

sequencing although cost effective, often requires batch processing of hundreds of microbial isolates to achieve cost savings and therefore this approach offers less flexibility for urgent, small scale sequencing often required during public health emergencies [10]. In contrast, Oxford Nanopore Technologies (ONT) offers a range of rapid real-time sequencing platforms from the portable MinION to the higher throughput GridION and PromethION models, although at this time lower read accuracy compared to Illumina data suggests accurate variant calling maybe problematic.

In September 2017, Public Health England (PHE) was notified of two cases of HUS in two children admitted to the same hospital on the same night. STEC O157:H7 was isolated from the faecal specimens of both cases. In order to rapidly determine whether or not the cases were part of a related phylogenetic cluster and therefore likely to be epidemiologically linked to each other, or to any other cases in the PHE database, we sequenced both isolates using the MinION platform and integrated the ONT sequencing data with a dense reference database of Illumina sequences. We aimed to compare Illumina and ONT sequencing data from the two isolates to assess the utility of the ONT method for urgent, small scale sequencing, and to determine whether the same single nucleotide variants were identified and whether inference of relatedness was consistent with the two technologies.

Data description

Paired-end FASTQ files were generated from the Illumina HiSeq 2500 for both samples (cases). Raw long-read data (FAST5) was generated from the MinION and basecalled using Albacore (FASTQ) in real-time. Both technologies derived FASTQ reads were trimmed and filtered (Trimmomatic, Porechop, Filtlong) before being aligned (BWA, Minimap2) to a reference genome (NC_002695.1). Variant positions were called using GATK before being imported into SnapperDB. Full processing details can be found within the methods section.

Results

Comparison of typing results generated by Illumina and ONT workflows

To consider the potential benefits of real-time sequencing to enhance opportunities for early outbreak detection, the timelines from DNA extraction to result generation for Illumina and ONT workflows were evaluated (Figure 1) and the relationship between yield, time and genome coverage plotted (Figure 2). For the ONT workflow, the time from DNA extraction to completion of the sequencing run was 28 hours. A total yield of 0.45 Gbases for the isolate from Case A and 0.59 Gbases for the isolate from Case B was achieved which corresponds to an equivalent coverage of the

Sakai O157 STEC reference genome (5.4Mb) of 81.29X and 108.30X for isolate A and B respectively. The average PHRED quality score for all reads in Case A was 9.87 and Case B was 9.47, which is approximately 1 error every 10 bases. Base-calling and analysis was performed in real-time and serotyping, Shiga toxin subtyping and variant identification were available within six hours and twenty minutes of the 24-hour sequencing run. With respect to the Illumina sequencing workflow, the time from DNA extraction to availability of results, assuming there were no breaks in the process, was just under 40 hours (Figure 1).

The species identification, serotype, MLST profile and Shiga toxin subtype results generated by both Illumina and ONT workflows were concordant with both isolates identified as *Escherichia coli* O157:H7 ST11, *stx2a* and *stx2c.* During the ONT sequencing run, the bacterial species was unambiguously identified in less than one minute for both cases (Figure 1). Additionally, using Krocus, a confirmed MLST was generated for Case A at 1:54 hours and Case B at 10:39 hours into the sequencing run. This was the point at which the last read required to generate a consensus on the MLST was base-called. By 93 minutes for Case A and 41 min for Case B, it was possible to determine the *E. coli* O157:H7 serotype, and *stx*2a and *stx*2c were detected at 58 and 24 minutes into the sequencing run for Case A and Case B, respectively.

Optimisation of ONT variant calling

To compare Illumina and ONT sequences within a standardised framework it was necessary to optimise the parameters for variant filtering within GATK2 to compensate for the lower read accuracy observed in the ONT data. Using Case B for the optimisation, base calls in the ONT data were classified as true positives (variant base detected by both methods), false positives (variant base in ONT, reference base in Illumina), true negatives (reference base in Illumina and ONT) or false negatives (variant base in Illumina, reference base in ONT). To disregard areas of the genome that the ONT reads could map to (and therefore identify variants) but were ambiguously mapped with Illumina reads, pre-filtering was performed by masking regions annotated as phage in the reference genome and those that could not be accurately self-mapped with simulated reference Illumina FASTQ reads. Figure 3 plots the precision (the proportion of true positives with respect to all positives calls) against the recall/sensitivity (the proportion of true positives identified with respect to all true positives) for an array of consensus ratio cut-offs for each of the masking strategies. Similar areas 'under the curve' were achieved for the different masking strategies with slightly higher precision at lower recall achieved with 'self-masking' (AUC – 0.71) and slightly higher recall at lower precision with explicit masking of the Sakai prophage (AUC – 0.75). The absence of a masking

strategy markedly affects the precision of variant calling with ONT data, in comparison of Illumina as a gold standard (AUC – 0.30). To identify the optimum consensus cut-off for filtering ONT variants processed through GATK the F1 score was calculated at each consensus cut-off. A consensus cut-off of 0.8 maximised the precision and recall (Figure 4) irrespective of the filtering methods.

Investigation of the discrepant variants identified between the Illumina and ONT data

After optimised quality and prophage filtering there were 266 and 101 base positions for Cases A and B respectively that were discordant between the ONT and Illumina sequencing data. The majority of discrepancies were where the ONT data identified a variant not identified in the Illumina data (261/266 (98.12%) and 95/101 (94.06%) discrepant base positions for Cases A and B respectively). In contrast the Illumina data identified 5 (1.88%) discrepant base positions as variants for Case A and 6 (5.94%) for case B (Table 1) not identified by the ONT data.

Variants and reason for omission.	**Case A**		**Case B**	
Total # of variants against the reference genome post quality filtering.	2076		1424	
Total # of variants with masked due to location in phage	708		531	
Total # of discrepant variants called between case A and B alone.	266		101	
Variants and reason for omission.	**Illumina VCF**	**ONT VCF**	**Illumina VCF**	**ONT VCF**
# of discrepant variants in each VCF.	5	261	6	95
# of discrepant variants with methylated positions masked.	0	260	0	94
Final discrepant variants.	5	1	6	1

Table 1 – Table showing the breakdown of the total number of variants of each technology against the reference genome, followed by the numbers of masked variants within prophage or methylated positions.

For both cases the most common discrepant variant were adenines classified as guanines in the ONT data with respect to the Illumina data (and reference), accounting for 68.05% (181/266) for Case A and 72.28% (73/101) for Case B. The second most common discrepancy was thymine being classified as cytosine in the ONT data accounting for 29.70% (79/266) in Case A and 22.80% (21/101) in Case B (Table 1). Of the transitions described above, 97.74% (Case A) and 93.07% (Case B) occurred when the variant was between two homopolymeric regions of multiple cytosines and guanines (Figure 5). These homopolymeric regions were similar to described DNA cytosine methylase (Dcm) binding sequences [11]. Nanopolish was subsequently used to identify likely Dcm, 5’ – cytosine – phosphate – guanine – 3’ (CpG) and DNA adenine methyltransferase (Dam) methylation sites in the ONT sequencing data and confirmed 260/266 (97.74%) and 94/101 (93.07%)

discrepant variants in the ONT data were classed as methylated for Cases A and B respectively. All of which were determined to be Dcm methylation for both cases.

Once the methylated positions were masked from the analysis, there were a total of 6 (5 discrepant variants in Illumina and 1 ONT) and 7 (6 discrepant variants in Illumina and 1 ONT) discrepant SNPs between the ONT and Illumina data, for Cases A and B respectively (Table 2 & 3). Four discrepant Illumina variants are shared by both Case A and Case B. One shared variant was found in a non-coding region, another shared variant was found in *rhsC* encoding an RHS (rearrangement hotspot) protein defined by the presence of extended repeat regions. Two further shared variants were found in in *dadX,* an alanine racemase gene. *dadX* is a paralogue of *alr*, also annotated as an alanine racemase in the *Sakai* reference genome with significant nucleotide similarity (>75% nucleotide identity). Both intra and inter gene repeats are known to be regions of potential false positives calls with Illumina data due to miss-mapping. Of the 7 variants in the Illumina data found in either or both Case A and B, 5 were found to be homoplastic in the O157 population of 4475 illumina sequences, arising independently, multiple times.

SNP	Position	BASE in Ref	BASE in Illum	Depth in Illum	BASE in ONT	Depth in ONT	Variant	Locus tag	Annotation
1	270,595	C	A	46	C	141	A	ECs0237	rhsC
2	379,516	A	G	114	A	100	G	NON CODING	N
3	1,681,338	C	G	59	C	61	G	ECs1685	alanine racemase 2
4	1,681,339	G	C	57	G	61	C	ECs1685	alanine racemase 2
5	2,636,513	T	C	91	T	69	C	ECs2674	hypothetical protein
6	4,709,195	A	A	86	G	82	G	ECs4673	membrane-bound ATP synthase epsilon-subunit AtpC

Table 2 – Table showing the final discrepant SNPs between the Illumina data and ONT data for case A. Also shown is the base as it is in the reference, the Ilumina called base and read depth at that position and the same for the ONT data. Finally, also included is the locus tag relative to the reference genome and the gene annotation.

SNP	Position	BASE in Ref	BASE in Illum	Depth in Illum	BASE in ONT	Depth in ONT	Variant	Locus tag	Annotation
1	270,595	C	A	19	C	207	A	ECs0237	rhsC
2	379,516	A	G	52	A	124	G	NON CODING	N
3	1,681,338	C	G	44	C	86	G	ECs1685	alanine racemase 2
4	1,681,339	G	C	41	G	86	C	ECs1685	alanine racemase 2
5	2,033,176	T	G	34	T	85	G	ECs2049	hypothetical protein
6	2,731,621	A	C	52	A	73	C	NON CODING	N

7	4,901,209	A	A	49	G	102	G	ECs4834	superoxide dismutase SodA

Table 3 – Table showing the final discrepant SNPs between the Illumina data and ONT data for case B. Also shown is the base as it is in the reference, the Ilumina called base and read depth at that position and the same for the ONT data. Finally, also included is the locus tag relative to the reference genome and the gene annotation.

Phylogenetic Analysis

Using the optimised variant calling parameters both strains clustered phylogenetically in lineage Ic within a dense reference database of STEC O157:H7 genomes (n=4475). However, the genomes were located in distinct sub-clades (Figure 6). It was, therefore, unlikely that the isolates originated from the same source, and it was concluded that Cases A and B were not epidemiologically linked. Following phylogenetic analysis of the Illumina SNP typing data (Figure 6), Case A was designated a sporadic case. However, Case B clustered with a concurrent outbreak, already under investigation, comprising three additional cases. The Illumina sequence linked to Case B was zero SNPs different from the other three cases in the cluster, whereas the ONT sequence was 7 SNPs different, when excluding the methylated positions (Table 3). Based on the ONT sequencing data alone, this discrepancy would have led to uncertainty as to whether or not the Case B was linked to the outbreak.

Assembly Profile

The ONT-only assembly resolved to five contigs (5.73 mb) for Case A and four contigs (5.60 mb) for Case B (Supplementary Table 1). In Case A, the five contigs were determined to be a single chromosomal contig, a single plasmid contig (pO157) and the three prophage duplications. In Case B, the four contigs were determined to be a single chromosomal contig with two plasmids (one being the pO157). For Case A the assembly resolved to 25 contigs (5.51mb) with a hybrid assembly and 668 contigs (5.45 mb) with an Illumina only assembly. Case B resolved to 34 contigs (5.49 mb) with a hybrid assembly and 575 contigs (5.42 mb) with an Illumina only assembly.

Alignment of the assemblies (Supplementary Figures 1 and 2) revealed several locations within the ONT-only assembly that there were absent in the hybrid and Illumina-only assemblies. In Case A, there were 8 regions only present within the ONT-only chromosome assembly, of which 7 are related to prophage regions (Supplementary Figure 1). In case B, there were 10 chromosomal regions in the ONT-only assembly that did not align to the other assemblies. All 10 regions were associated with prophage regions (Supplementary Figures 2).

Discussion

In this study, the two isolates sequenced using ONT were unambiguously identified as STEC O157:H7 ST 11 *stx2a*/*stx2c* in less than 15 hours and it was possible to distinguish the genetic relatedness between the isolates within 377 minutes (i.e. 22 hours before the ONT sequencing run was scheduled to finish and just under three hours before the Illumina sequencing began.). The WGS turn-around time from DNA extraction and library preparation, to sequencing and analysis via the Illumina workflow at PHE, is three to six days. Although this turnaround time is rapid for a service utilising batch processing on the HiSeq platforms, the sequencing approach using the MinION, whereby individual samples or small barcoded batches are loaded and results generated and analysed in real-time, has the potential to be faster and more flexible. It should be noted that speed to result for ONT sequencing will be related to the amount of isolates run on a single flowcell as DNA molecules from different samples will compete to traverse a finite number of pores. This approach is therefore ideal for urgent, small scale sequencing, often required during public health emergencies. In this scenario, analysis of the ONT data provided evidence that the two cases were not epidemiologically linked and, although efforts were made to determine the potential source of the infection for both cases through the National Enhanced STEC Surveillance System [2], an outbreak investigation was not initiated.

A current limitation of MinION sequencing is its lower read accuracy when compared to short-read technologies [12,13,14,15,16]. This accuracy has improved as the technology has matured but still falls short of the 99% accuracy offered by short-read platforms [15]. There are a number of factors that contribute to the current read accuracy in the nanopore data including structural similarity of nucleotides, simultaneous influence of multiple nucleotides on the signal, the non-uniform speed at which nucleotides pass through the pore and the fact that the signal does not change within homoploymers [15].

Although analysis of the Illumina and ONT sequencing data placed the sequences on the same branch on the phylogeny, there were SNP discrepancies between the sequences generated by the two different workflows, even after optimisation of the parameters. The vast majority of the discrepant SNPs (261/266 – 98.12% and 95/101 – 94.06 % for Cases A and B respectively) were attributed to variants identified in the ONT data and not the Illumina data. The majority of discrepancies (97.74% in Case A and 93.07% in Case B) were found in sequences that are the same as the known 5-Methylcytosine motif sequences, CC(A/T)GG [11,17] in the ONT data. Following a

search of the ONT discrepant SNPs for CpG, Dam and Dcm methylation using Nanopolish, the majority (97.74% and 93.07% for case A and B respectively) of the ONT discrepant SNPs were identified in Dcm methylated regions.

As Nanopolish is detecting these methylated positions with the use of the raw FAST5 data, it is suggested that these particular discrepancies appear during the basecalling process. Albacore handles most methylation well across the three methylation models searched for by Nanopolish, for example only 94 out of 13,504 methylated positions were considered incorrect by base calling for Case B. However, for mapping based-SNP typing, this level of error in base calling means that it is not possible to accurately determine the number of SNPs, thus potentially obscuring the true phylogenetic relationship between isolates of STEC O157:H7.

The optimisation of variant filtering was performed using the Illumina data as a gold standard. However, it is possible that the alignment of the Illumina data might have generated false SNPs based on reads mapping to ambiguous regions of the genome, whereas the long reads obtained using the ONT workflow are able to resolve these ambiguous regions and call variants, or not, at these positions correctly. As the Illumina data was used as the gold standard, in this scenario SNPs produced in the Illumina data would have been classed incorrectly as false negatives in the ONT data. Discrepant variants identified in the Illumina data were attributed mainly to potentially false mapping of Illumina reads to homologous regions of the reference genome, variants which were misidentified at the same position independently in Case A and Case B. Furthermore, comparison of assemblies generated by ONT reads, Illumina reads and a hybrid approach highlights the extra genetic content accessible to ONT assemblies where variation can be quantified.

In this study an ONT sequencing workflow was used to rapidly rule out an epidemiological link between two children admitted to the same hospital on the same day with symptoms of HUS. The isolates of STEC O157:H7 from each child mapped to different clades within the same STEC O157:H7 lineage (Ic). We provide further evidence that SNP typing using MinION-based variant calling is possible when the coverage of the variation is high [15]. The error rate exhibited by ONT sequencing workflows continue to improve due to developments in the pore design, the library preparation methods, innovations in base-calling algorithms and the introduction of post-sequencing correction tools, such as Nanopolish [15,21]. Currently, both short and long read technologies are used for public health surveillance, and there is a need to integrate the outputs so that all the data can be analysed in the same way. Recently, Rang et al [15] reiterated how the scientific community can

make valuable contributions to improving ONT read accuracy by systematically comparing computational strategies as highlighted in this study and elsewhere [22]. On-going updates to the chemistry and software tools will facilitate the robust detection of SNPs enabling ONT to compete with short read platforms, ultimately enabling the two technologies to be used interchangeably in clinical and public health settings.

Methods

DNA extraction, Library preparation and Illumina Sequencing

Genomic DNA was extracted from two strains of STEC O157 isolates from two HUS cases admitted to the same hospital on the same night. Using a Qiagen Qiasymphony (Qiagen, Hilden, Germany) to manufacture's instructions, genomic DNA extracted and quantified using a Qubit and the BR dsDNA Assay Kit (Thermofisher Scientific, Waltham, USA) to manufacture's instructions. The sequencing library was prepared by fragmenting and tagging the purified gDNA using the Nextera XT DNA Sample Preparation Kits (Illumina, Cambridge, UK) to manufacture's instructions. The prepared library was loaded onto an Illumina HiSeq 2500 (Illumina, Cambridge, UK) at PHE and sequencing perfomed in rapid run mode yielding paired-end 100bp reads.

Processing and analysis of Illumina sequence data

FASTQ reads were processed using Trimmomatic v0.27 (Trimmomatic , RRID:SCR_011848)[23] to remove bases with a PHRED score of less than 30 from the leading and trailing ends, with reads less than 50 bp after quality trimming discarded. A *k*-mer approach (https://github.com/phe-bioinformatics/kmerid) was used to confirm the species of the samples. Sequence type (ST) assignment was performed using MOST v1.0 described by [24]. *In silico* serotyping was performed by using GeneFinder (https://github.com/phe-bioinformatics/gene_finder) which uses Bowtie v2.2.5 (Bowtie , RRID:SCR_005476)[25] and Samtools v0.1.18 (SAMTOOLS , RRID:SCR_002105)[26] to align FASTQ reads to a multifasta containing the target genes (including *wzx*, *wzy* and *fliC*). *Stx* sub-typing was performed as described in [27]. Illumina FASTQ reads were mapped to the Sakai STEC O157 reference genome (NC_002695.1) using BWA MEM v0.7.13 (BWA , RRID:SCR_010910)[28]. Variant positions identified by GATK v2.6.5 UnifiedGenotyper (GATK , RRID:SCR_001876)[29] that passed the following parameters; >90% consensus, minimum read depth of 10, Mapping Quality (MQ) >= 30. Any variants called at positions that were within the known prophages in Sakai were masked from further analyses. The remaining variants were imported into SnapperDB v0.2.5 [30].

DNA extraction, Library preparation and Nanopore Sequencing

To preserve DNA integrity for the nanopore sequencing, genomic DNA was extracted and purified using the Promega Wizard Genomic DNA Purification Kit (Promega, Madison, USA) with minor alterations including doubled incubation times, no vigorous mixing steps (performed by inversion) and elution into 50µl of double processed nuclease free water (Sigma-Aldrich, St. Louis, USA). DNA was quantified using a Qubit and the HS (High sensitivity) dsDNA Assay Kit (Thermofisher Scientific, Waltham, USA) to manufacture's instructions. Library preparation was performed using the Rapid Barcoding Kit - SQK-RBK001 (Oxford Nanopore Technologies, Oxford, UK) with each sample's gDNA being barcoded by transposase based tagmentation and pooled as per manufacture's instructions. The prepared library was loaded on a FLO-MIN106 R9.4 flow cell (Oxford Nanopore Technologies, Oxford, UK) and sequenced using the MinION for 24 hours.

Processing and analysis of Nanopore sequence data

Raw FAST5 files were basecalled and de-multiplexed in real-time, as reads were being generated, using Albacore v2.1 (Oxford Nanopore Technologies) into FASTQ files. Run metrics were generated using Nanoplot v1.8.1 using default parameters [31]. Reads were processed through Porechop v0.2.1 using default parameters (Wick. Unpublished) [32] to remove any barcodes and adapters used in SQK-RBK001. For Case A 96,788 reads (10,214,353 bases) were adaptor trimmed and 386 (0.39%) chimeric reads split. For Case B 430,911 reads (34,888,999 bases) were adaptor trimmed and 513 (0.11%) chimeric reads split. Samples were speciated using Kraken v0.10.4 [33]. A MLST was assigned using Krocus with the following parameters --kmer 15, --min_block_size 300 and --margin 500 [34]. S*tx* sub-typing and serotyping was determined by aligning the basecalled reads using minimap2 v2.2 [35] and Samtools v1.1 [26] to a multifasta containing the *Stx* and serotype encoding genes.

For reference based variant calling FASTQ reads were mapped to the Sakai STEC O157 reference genome (NC_002695.1) using minimap2 v2.2 [35]. VCFs were produced using GATK v2.6.5 UnifiedGenotyper [29]. Any variants called at positions that were within the known prophages in Sakai were masked from further analyses. To determine the optimum consensus cut-off for ONT variant detection the VCF was filtered with sequentially decreasing ad-ratio values at 0.1 intervals. Using the Illumina variant calls as the gold standard, F1 scores (the weighted average of precision and recall) were calculated to determine the optimal ad-ratio for processing ONT data through GATK.

Comparison of Illumina and Nanopore discrepant SNPs

Nanopolish [21] was also used to detect methylation across the ONT data to compare to the discrepant positions. This was performed using the call-methylation function searching for three types of methylation including, the DNA adenine methyltransferase (Dam), DNA cytosine methylase (Dcm) and 5' – cytosine – phosphate – guanine – 3' (CpG) models. The discrepant SNPs between the Illumina and ONT for both Case A and Case B were manually visualised in Tablet v1.17.08.17 [36] in order to elucidate the reason for the discrepancy. Discordant SNPs being within a homopolymeric region were also quantified.

Generation of phylogenetic trees

Filtered VCF files for each of the Illumina and ONT sequencing data for each sample, were incorporated, into SnapperDB v0.2.5 [30] containing variant calls from 4471 other STEC CC11 genomes generated through routine surveillance by Public Health England. SnapperDB v0.2.5 [30] was used to generate a whole genome alignment of the 4475 genomes (including both datasets for the selected strains for this study). Both methylated positions and prophage positions were masked from the alignment. The alignment was processed through Gubbins V2.0.0 [37] to account for recombination events. A maximum likelihood tree was then constructed using RAxML V8.1.17 [38].

Assembly of ONT data

Trimmed ONT FASTQ files were assembled using Canu v1.6 (Canu, RRID:SCR_015880)[39]. Polishing of the assemblies was performed using Nanopolish v0.10.2 [21] using both the trimmed ONT FASTQs and FAST5s for each respective sample accounting for methylation using the --methylation-aware option set to dcm. Assemblies were reoriented to start at the *dnaA* gene (NC_000913) from *E. coli* K12, using the fixstart parameter in circulator v1.5.5 [40].

Hybrid assemblies

Trimmed ONT FASTQ files were assembled using Unicycler v0.4.2 [41] with the following parameters min_fasta_length=1000, mode=normal and -1 and -2 for the incorporation of each sample's equivalent Illumina FASTQ. Pilon v1.23 [42] was used to correct the assembly using the Illumina reads.

Assembly of Illumina data

Illumina reads were assembled using SPAdes v3.13.0 (SPAdes , RRID:SCR_000131)[43] with the careful parameter activated and with kmer lengths of 21, 33, 55, 65, 77, 83 and 91.

Annotation

Prokka v1.13 [44] with the species set to E. coli was used to annotate the final assemblies. Mauve snapshot_2015-02-25 (1) [45] using the "move contig" function was used to align each assembly to the ONT reference as they had the least number of contigs.

Availability of supporting data

The FASTQ files for the paired read Illumina sequence data can be found on the NCBI (National Center for Biotechnology Information) Sequence Read Archive (SRA); Case A accession: SRR7184397, Case B accession - SRR6052929. The ONT FASTQ files, Case A accession – SRR7477814, Case B accession - SRR7477813. All files can be found under BioProject - PRJNA315192. The assemblies in the supplementary data were submitted to NCBI GenBank. Illumina assemblies, Case A – X and Case B – X. ONT assemblies, Case A – X and Case B. All supplementary files can be found under BioProject - PRJNA315192, and additional supporting data is in the *GigaScience* GigaDB respository [46].

Abbreviations

AUC: Area under curve; BWA: Burrows-Wheeler aligner; CC: Clonal complex; Dam: DNA adenine methyltransferase; Dcm: DNA cytosine methylase; GATK: Genome analysis toolkit; HUS: Haemolytic Uremic Syndrome; MLST: Multi-locus sequence type; NCBI: National Center for Biotechnology Information; ONT: Oxford Nanopore Technologies; PHE: Public Health England; SNP: Single nucleotide polymorphism; SRA: Sequence read archive; STEC: Shiga toxin-producing *Escherichia coli*; VCF: Variant call format; WGS: Whole genome sequencing.

Author contributions

CJ and TJD conceptualised the project. CJ and DRG performed DNA extractions. DRG performed library preparation and Nanopore sequencing. TJD and DRG processed Illumina sequence data. DRG processed all ONT data. TJD performed ONT optimisation. DRG performed methylation analysis. TJD and DRG performed Illumina and ONT data comparison. CJ wrote the original draft. DRG, CJ, SG and TJD performed manuscript editing.

Competing interests

This project was part funded by Oxford Nanopore Technologies.

Acknowledgements

We would like to thank Oxford Nanopore Technologies for supplying the Rapid Barcoding Kit - SQK-RBK001 and FLO-MIN106 R9.4 flow cell used in this research. In particular we would like to thank Leila Luheshi and Divya Mirrington for their support and scientific assistance.
We would also like to thank the frontline NHS Laboratories for submitting the samples used in this study to the Gastrointestinal Bacteria Reference Unit at Public Health England.
We would like to acknowledge Dr Andrew Page of the Quadram Institute for critically reviewing the manuscript.

References

1. Croxen MA, Law RJ, Scholz R, Keeney KM, Wlodarska M, Finlay BB. Recent advances in understanding enteric pathogenic *Escherichia coli*. Clin Microbiol Rev. 2013. 26(4):822-80. doi: 10.1128/CMR.00022-13.

2. Byrne L, Jenkins C, Launders N, Elson R, Adak GK. The epidemiology, microbiology and clinical impact of Shiga toxin-producing *Escherichia coli* in England, 2009-2012. Epidemiol Infect. 2015. 143(16):3475-87. doi: 10.1017/S0950268815000746.

3. Launders N, Byrne L, Jenkins C, Harker K, Charlett A, Adak GK. Disease severity of Shiga toxin-producing *E. coli* O157 and factors influencing the development of typical haemolytic uraemic syndrome: a retrospective cohort study, 2009-2012. BMJ Open. 2016. 6(1):e009933. doi: 10.1136/bmjopen-2015-009933.

4. Heiman KE, Mody RK, Johnson SD, Griffin PM, Gould LH. *Escherichia coli* O157 Outbreaks in the United States, 2003-2012. Emerg Infect Dis. 2015. 21(8):1293-1301. doi: 10.3201/eid2108.141364.

5. Dallman TJ, Ashton PM, Byrne L, Perry NT, Petrovska L, Ellis R, Allison L, Hanson M, Holmes A, Gunn GJ, Chase-Topping ME, Woolhouse ME, Grant KA, Gally DL, Wain J, Jenkins C. Applying phylogenomics to understand the emergence of Shiga-toxin-producing *Escherichia coli* O157:H7 strains causing severe human disease in the UK. Microb Genom. 2015. 1(3):e000029. doi: 10.1099/mgen.0.000029.

6. 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. Whole-genome sequencing for national surveillance of Shiga toxin-producing *Escherichia coli* O157. Clin Infect Dis. 2015. 61(3):305-12. doi: 10.1093/cid/civ318.

7. Olson ND, Lund SP, Colman RE, Foster JT, Sahl JW, Schupp JM, Keim P, Morrow JB, Salit ML, Zook JM. Best practices for evaluating single nucleotide variant calling methods for microbial genomics. Front Genet. 2015. 6(235): doi: 10.3389/fgene.2015.00235.

8. Ruffalo M, Koytürk M, Ray S, LaFramboise T. Accurate estimation of short read mapping quality for next-generation genome sequencing. Bioinformatics. 2012. 28(18):i349-i355. doi: 10.1093/bioinformatics/bts408.

9. Timme RE, Rand H, Sanchez Leon M, Hoffmann M, Strain E, Allard M, Roberson D, Baugher JD. GenomeTrakr proficiency testing for foodborne pathogen surveillance: an exercise from 2015. Microb Genom. 2018. 4(7). doi: 10.1099/mgen.0.000185.

10. Quick J, Loman NJ, Duraffour S, Simpson JT, Severi E, Cowley L, Bore JA, Koundouno R, Dudas G, Mikhail A, Ouédraogo N, Afrough B, Bah A, Baum JH, Becker-Ziaja B, Boettcher JP, Cabeza-Cabrerizo M, Camino-Sanchez A, Carter LL, Doerrbecker J, Enkirch T, Dorival IGG, Hetzelt N, Hinzmann J, Holm T, Kafetzopoulou LE, Koropogui M, Kosgey A, Kuisma E, Logue CH, Mazzarelli A, Meisel S, Mertens M, Michel J, Ngabo D, Nitzsche K, Pallash E, Patrono LV, Portmann J, Repits JG, Rickett NY, Sachse A, Singethan K, Vitoriano I, Yemanaberhan RL, Zekeng EG, Trina R, Bello A, Sall AA, Faye O, Faye O, Magassouba N, Williams CV, Amburgey V, Winona L, Davis E, Gerlach J, Washington F, Monteil V, Jourdain M, Bererd M, Camara A, Somlare H, Camara A, Gerard M, Bado G, Baillet B, Delaune D, Nebie KY, Diarra A, Savane Y, Pallawo RB, Gutierrez GJ, Milhano N, Roger I, Williams CJ, Yattara F, Lewandowski K, Taylor J, Rachwal P, Turner D, Pollakis G, Hiscox JA, Matthews DA, O'Shea MK, Johnston AM, Wilson D, Hutley E, Smit E, Di Caro A, Woelfel R, Stoecker K, Fleischmann E, Gabriel M, Weller SA, Koivogui L, Diallo B, Keita S, Rambaut A, Formenty P, Gunther S, Carroll MW. Real-time portable genome sequencing for Ebola surveillance. Nature. 2016. 530(7589):228-32 doi: 10.1038/nature16996.

11. Gomez-Eichelmann MC, Levy-Mustri A, Ramirez-Santos J. Presence of 5-methycytosine in CC(A/T)GG sequences (Dcm methylation) in DNAs from different bacteria. J Bacteriol. 1991. 173(23):7692-4.

12. Mikheyev AS, Tin MM. A first look at the Oxford Nanopore MinION sequencer. Mol Ecol Resour. 2014. 14(6):1097-102. doi: 10.1111/1755-0998.12324.

13. Laver T, Harrison J, O'Neill PA, Moore K, Farbos A, Paszkiewicz K, Studholme DJ. Assessing the performance of the Oxford Nanopore Technologies MinION. Biomol Detect Quantif. 2015. 1-8.

14. Magi A, Semeraro R, Mingrino A, Giusti B, D'Aurizio R. Nanopore sequencing data analysis: state of the art, applications and challenges. Brief Bioinform. 2017. doi: 10.1093/bib/bbx062.

15. Rang FJ, Kloosterman WP, de Ridder J. From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. Genome Biol. 2018. 19(1):90. doi: 10.1186/s13059-018-1462-9.

16. Senol Cali D, Kim JS, Ghose S, Alkan C, Mutlu O. Nanopore sequencing technology and tools for genome assembly: computational analysis of the current state, bottlenecks and future directions. Brief Bioinform. 2018. doi: 10.1093/bib/bby017.

17. Marinus MG. DNA methylation in *Escherichia coli*. Annu Rev Genet. 1987. 21:113-31.

18. Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, Tyson JR, Beggs AD, Dilthey AT, Fiddes IT, Malla S, Marriott H, Nieto T, O'Grady J, Olsen HE, Pedersen BS, Rhie A, Richardson H, Quinlan AR, Snutch TP, Tee L, Paten B, Phillippy AM, Simpson JT, Loman NJ, Loose M. Nanopore sequencing and assembly of a human genome with ultra-long reads. Nat Biotechnol. 2018. 36(4):338-345. doi: 10.1038/nbt.4060.

19. Ebler J, Haukness M, Pesout T, Marschall T, Paten B. Haplotype-aware diplotyping from noisy long reads. Genome Biol. 2019 Jun 3;20(1):116. doi: 10.1186/s13059-019-1709-0.

20. Sarkozy P, Jobbágy Á, Antal P. Calling homopolymer stretches from raw Nanopore reads by analyzing k-mer dwell times. In: Eskola H., Väisänen O., Viik J., Hyttinen J. (eds) EMBEC & NBC 2017. EMBEC 2017, NBC 2017. IFMBE Proceedings, 2018. vol 65. Springer, Singapore.

21. Loman NJ, Quick J, Simpson JT. A complete bacterial genome assembled de novo using only nanopore sequencing data. Nat Methods. 2015. 12(8):733–5. doi: 10.1038/nmeth.3444.

22. Wick RR, Judd LM, Gorrie CL, Holt KE. Completing bacterial genome assemblies with multiplex MinION sequencing. Microb Genom. 2017. 3(10):e000132. doi: 10.1099/mgen.0.000132.

23. Bolger AM, Lohse M, Usadel B. Trimmomatic: A flexible trimmer for Illumina Sequence Data. Bioinformatics. 2014. 30(15):2114-20. doi: 10.1093/bioinformatics/btu170.

24. Tewolde R, Dallman T, Schaefer U, Sheppard CL, Ashton P, Pichon B, Ellington M, Swift C, Green J, Underwood A. MOST: a modified MLST typing tool based on short read sequencing. PeerJ. 2016. 4:e2308. doi: 10.7717/peerj.2308.

25. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012. 9(4):357-9. doi: 10.1038/nmeth.1923.

26. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. The Sequence alignment/map (SAM) format and SAMtools. Bioinformatics. 2009. 25(16):2078-9. doi: 10.1093/bioinformatics/btp352.

27. Ashton PM, Perry N, Ellis R, Petrovska L, Wain J, Grant KA, Jenkins C, Dallman TJ. Insight into Shiga toxin gene encoded by *Escherichia coli* O157 from whole genome sequencing. PeerJ. 2015. 17. doi: 10.7717/peerj.739.

28. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler Transform. Bioinformatics. 2009. 25(14):1754-60. doi: 10.1093/bioinformatics/btp324.

29. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. The Genome Analysis Toolkit: a MapReduce frame- work for analyzing next-generation DNA sequencing data. Genome Res. 2010. 20(9):1297-303. doi: 10.1101/gr.107524.110.

30. Dallman T Ashton P, Schafer U, Jironkin A, Painset A, Shaaban S, Hartman H, Myers R, Underwood A, Jenkins C, Grant K. SnapperDB: A database solution for routine sequencing analysis of bacterial isolates. Bioinformatics. 2018. doi: 10.1093/bioinformatics/bty212.

31. De Coster W, D'Hert S, Schultz DT, Cruts M, Van Broeckhoven C. NanoPack: visualizing and processing long-read sequencing data. Bioinformatics. 2018. 34(15):2666-9. doi: 10.1093/bioinformatics/bty149.

32. Wick R. Porechop GitHub page. https://github.com/rrwick/Porechop.

33. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol. 2014. 15(3):R46. doi: 10.1186/gb-2014-15-3-r46.

34. Page A, Keane J. Rapid multi-locus sequence typing direct from uncorrected long reads using Krocus. PeerJ. 2018. 6:e5233 doi: 10.7717/peerj.5233.

35. Li H. Minimap2: fast pairwise alignment for long nucleotide sequences. Bioinformatics. 2018. doi: 10.1093/bioinformatics/bty191.

36. Milne I Stephen G, Bayer M, Cock PJ, Pritchard L, Cardle L, Shaw PD, Marshall D. Using Tablet for visual exploration of second-generation sequencing data. Briefings in Bioinformatics. 2013. 14(2):193-202. doi: 10.1093/bib/bbs012.

37. Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA, Bentley SD, Parkhill J, Harris SR. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. Nucleic Acids Res. 2014. 43(3):e15. doi: 10.1093/nar/gku1196.

38. Stamatakis A. 2014. RAxML Version 8: A tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies. Bioinformatics. 2014. 30(9):1312-13. doi: 10.1093/bioinformatics/btu033.

39. Koren S, Walenz BP, Berlin K, Miller JR, Phillippy AM. 2017. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. 27(5):722-36. doi: 10.1101/gr.215087.116.

40. Hunt M, Silva ND, Otto TD, Parkhill J, Keane JA, Harris SR. 2015. Circlator: automated circularization of genome assemblies using long sequencing reads. Genome Biol. 16(294):1-10. doi: 10.1186/s13059-015-0849-0.

41. Wick RR, Judd LM, Gorrie CL, Holt KE. 2017. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. PLoS Comput Biol. 13(6):e1005595. doi: 10.1371/journal.pcbi.1005595.

42. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, Earl AM. 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLOS One. 9(11):e112963. doi: 10.1371/journal.pone.0112963.

43. 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 sigle cell sequencing. J Comput Biol. 19(5):455-77. doi: 10.1089/cmb.2012.0021.

44. Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 30(14):2068-9. doi: 10.1093/bioinformatics/btu153.

45. Darling ACE, Mau B, Blattner FR, Perna NT. 2004. Mauve: Multiple alignment of conserved genomic sequence with rearrangements. Genome Res. 14(7):1394-403. doi: 10.1101/gr.2289704.

46. Greig DR; Jenkins C; Gharbia S; Dallman TJ (2019): Supporting data for "Comparison of single nucleotide variants identified by Illumina and Oxford Nanopore technologies in the context of a potential outbreak of Shiga Toxin Producing Escherichia coli" GigaScience Database. http://dx.doi.org/10.5524/100623

597

598

599

600

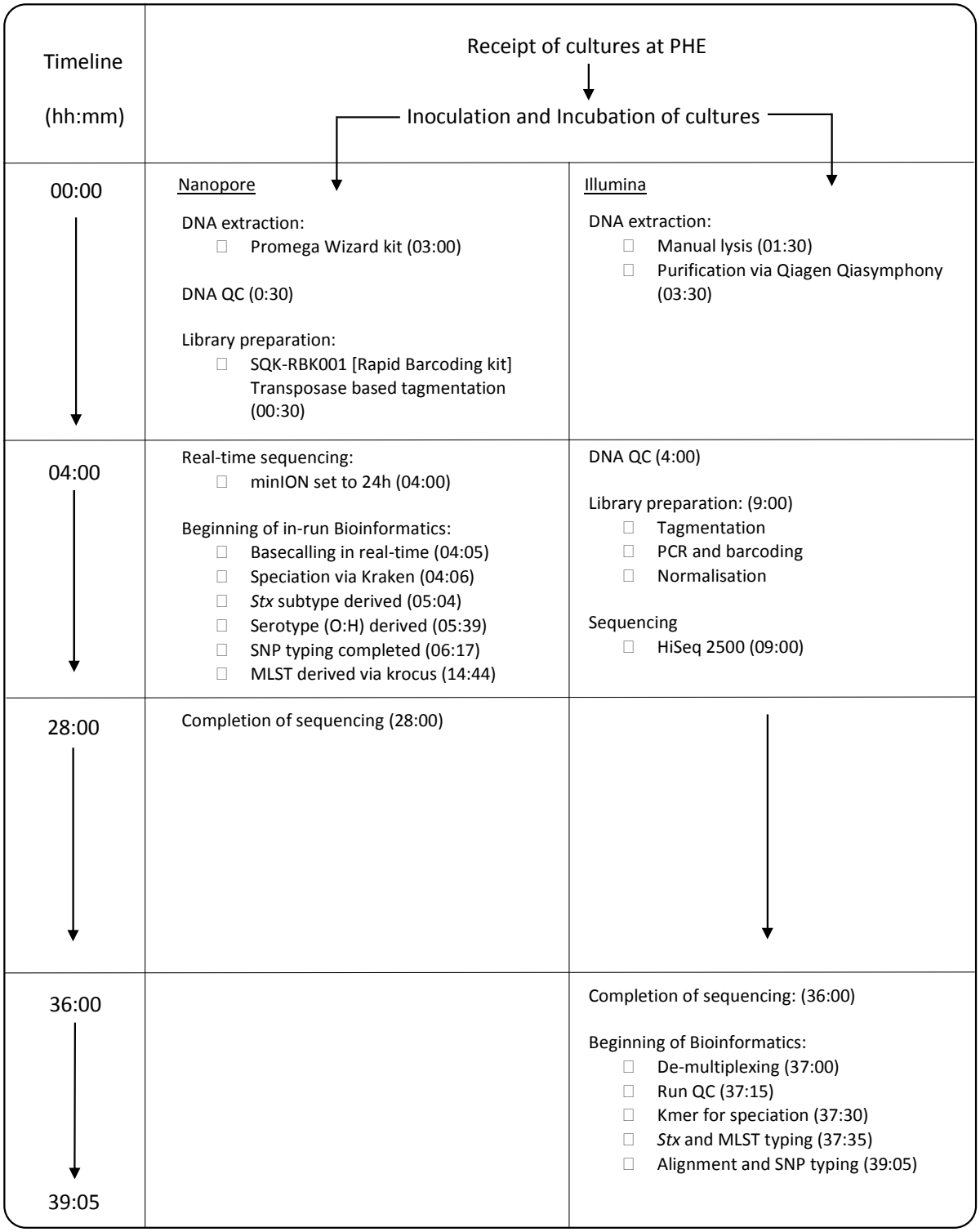


Figure 1 – Figure showing comparative timeline from beginning DNA extraction to results generation for Oxford Nanopore and Illumina technologies. Times shown the completion of the labelled event relative to the start of the assay (hh:mm).

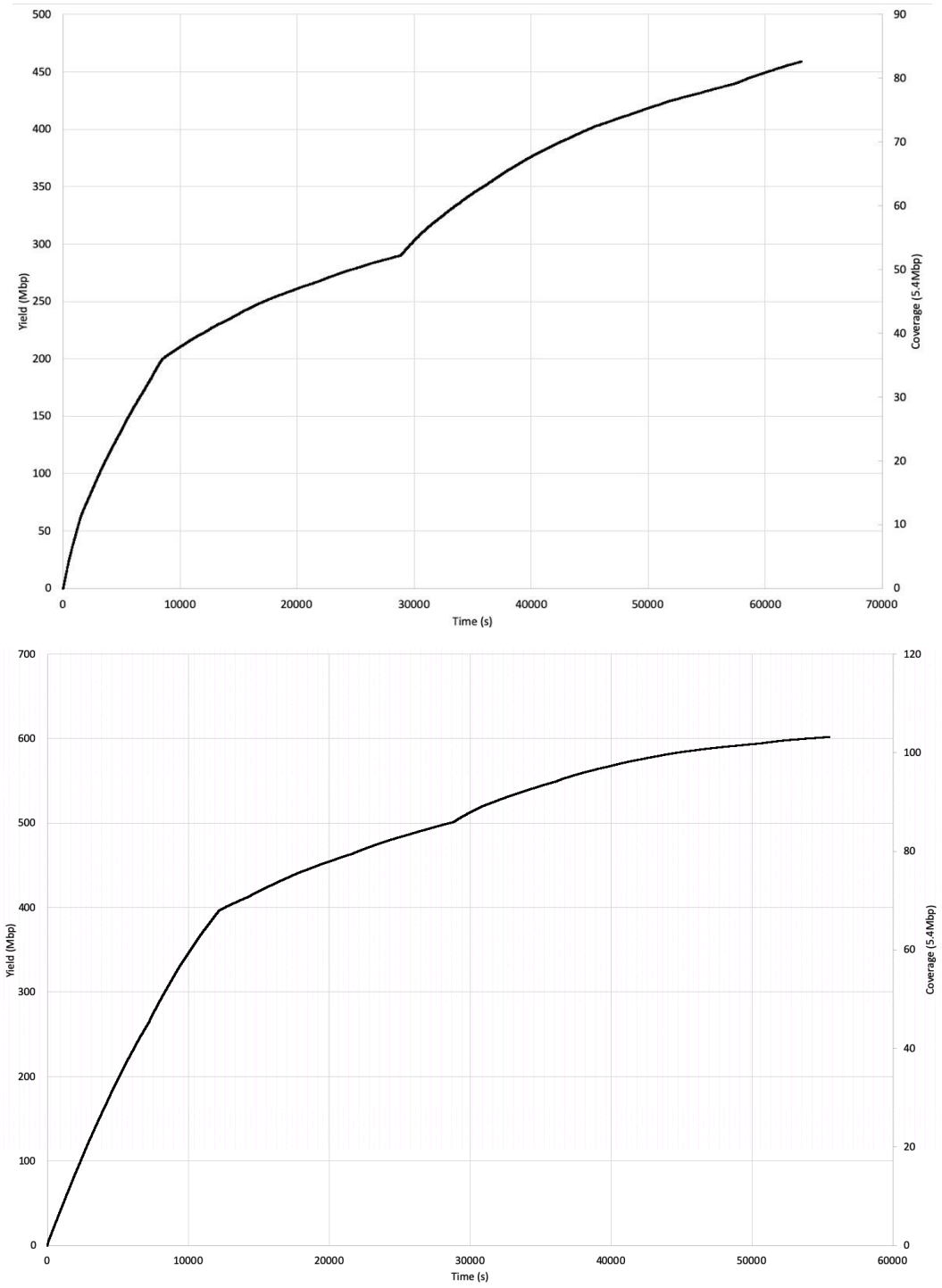


Figure 2 – Two time/yield/coverage graphs showing production of reads in real-time and the associated cumulative mapping coverage. Case A is the graph on the left and Case B is on the right.

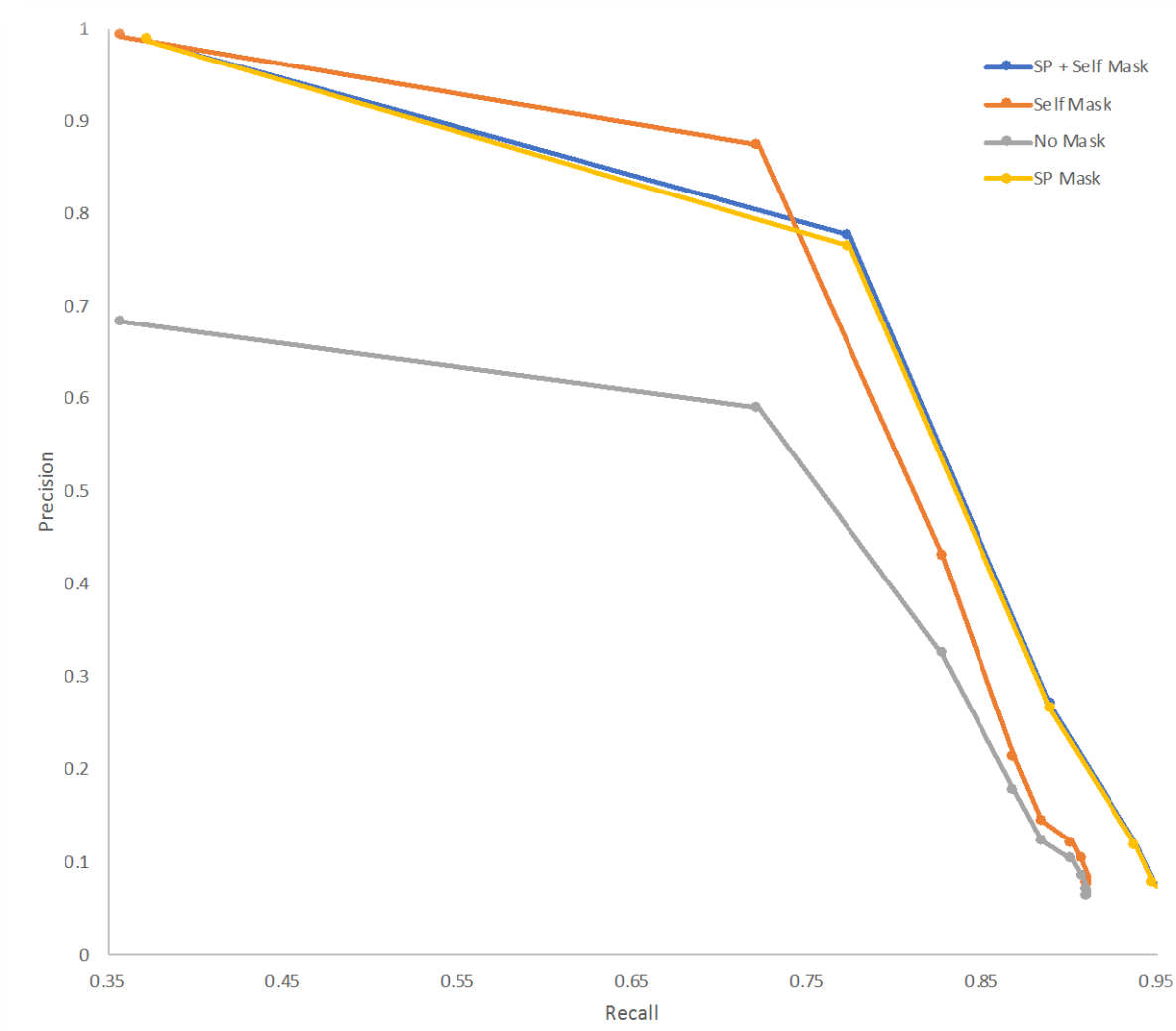


Figure 3 – Precision Vs Recall of variant calling for an array of consensus ratio cut-offs and pre-masking strategies including masking positions annotated as 'Sakai phage' ('SP') and positions that are ambiguously self-mapped ('Self') with simulated Illumina FASTQs from the reference genome. Performed on case B.

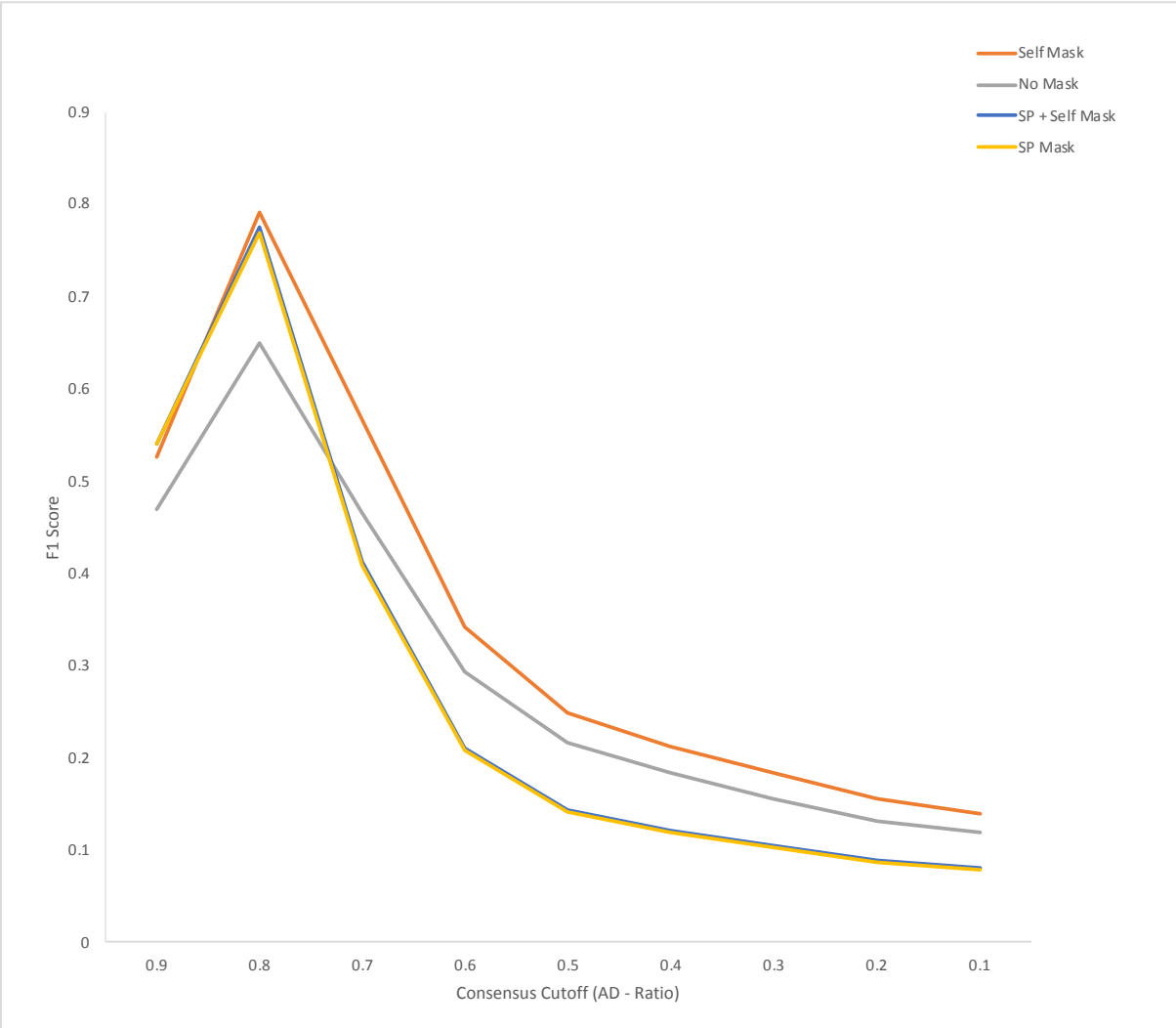


Figure 4 – F1 Score for an array of consensus ratio cut-offs and pre-masking strategies including masking positions annotated as 'Sakai phage' ('SP') and positions that are ambiguously self-mapped ('Self') with simulated Illumina FASTQs from the reference genome.

	Position						Case	
	-2	**-1**	**Variant**	**+1**	**+2**		**A**	**B**
Reference	C	C	A	G	G	A > G Transition	69.62% (n=181)	77.66% (n=73)
			↓					
Alignment	C	C	G	=	G			
Reference	C	C	T	G	G	T > C Transition	30.38% (n=79)	22.34% (n=21)
			↓					
Alignment	C	C	C	G	G			
						Total	100% (n=260)	100% (n=94)

Figure 5 – Figure showing the two most common discrepancies in the ONT optimised GATK VCFs and a breakdown of the relative proportions of these transitions compared to the total number of discrepant SNPs for both cases.

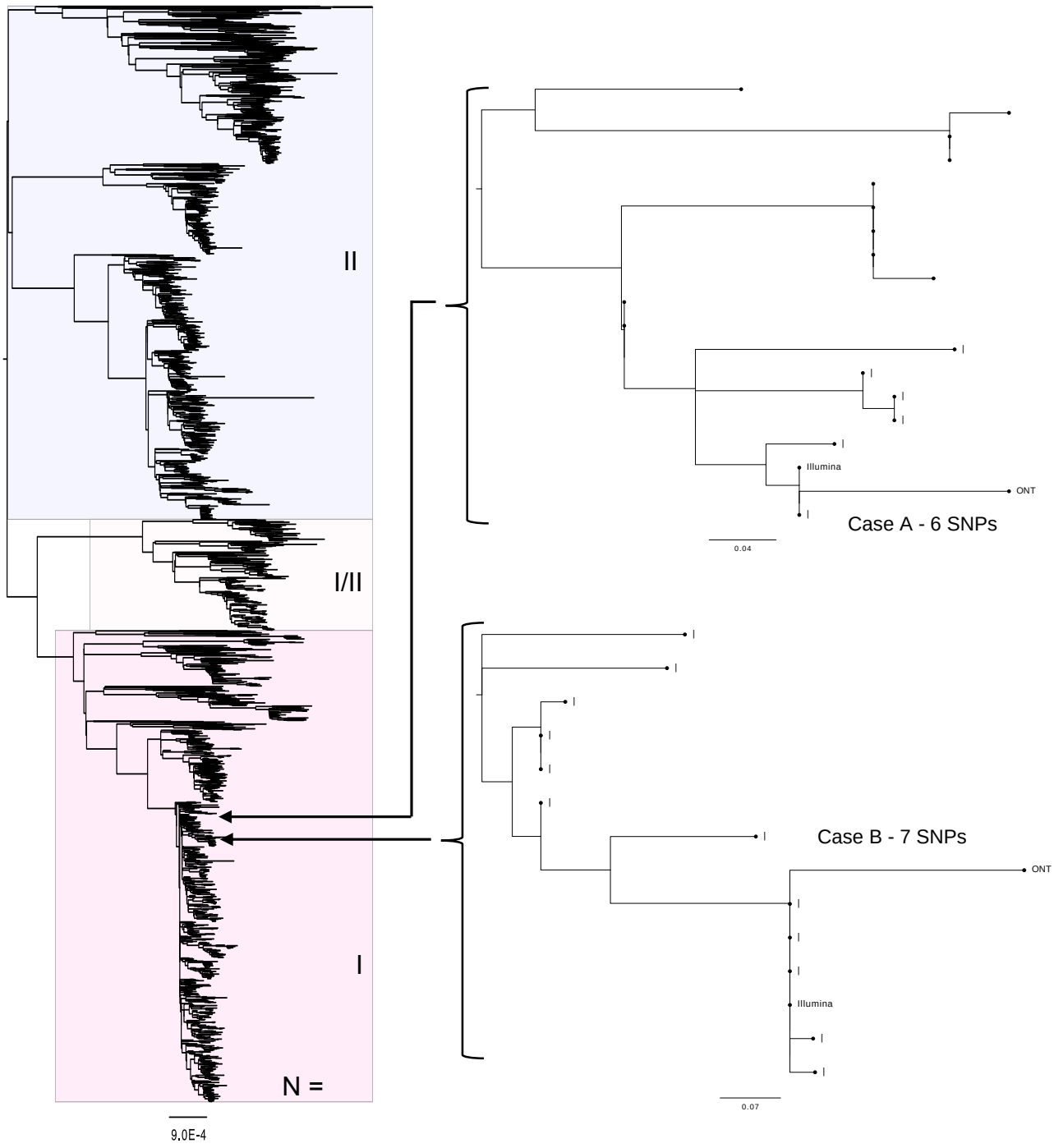


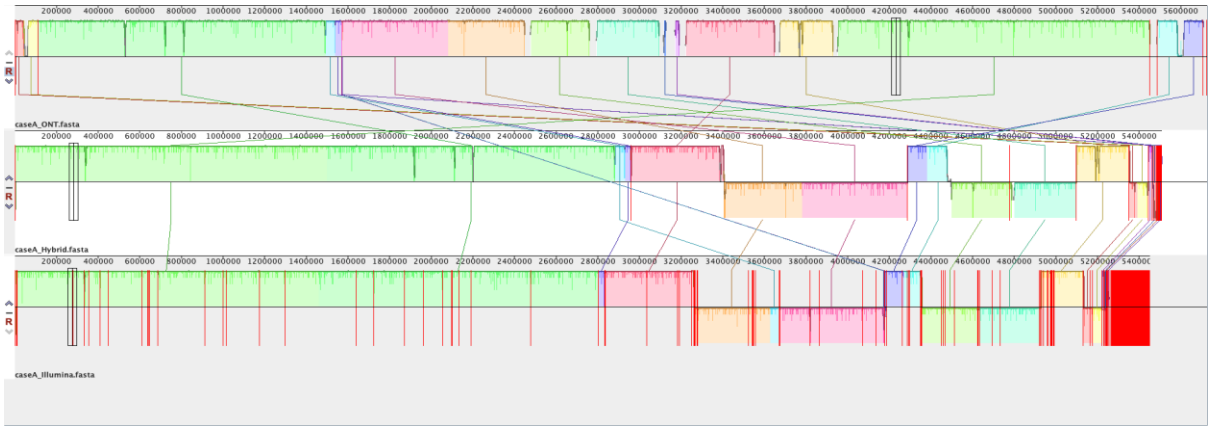
Figure 6 – Maximum likelihood tree, of a "soft core" alignment of 4475 genomes showing the tree lineages (I, I/II and II) of STEC (Clonal Complex 11). Also showing where Oxford Nanopore and Illumina sequencing data is placed within the tree for each of the two cases. All methylated positions and prophage regions have been masked. Values represent the SNP differences between the Illumina and ONT data for both cases.

Supplementary Tables

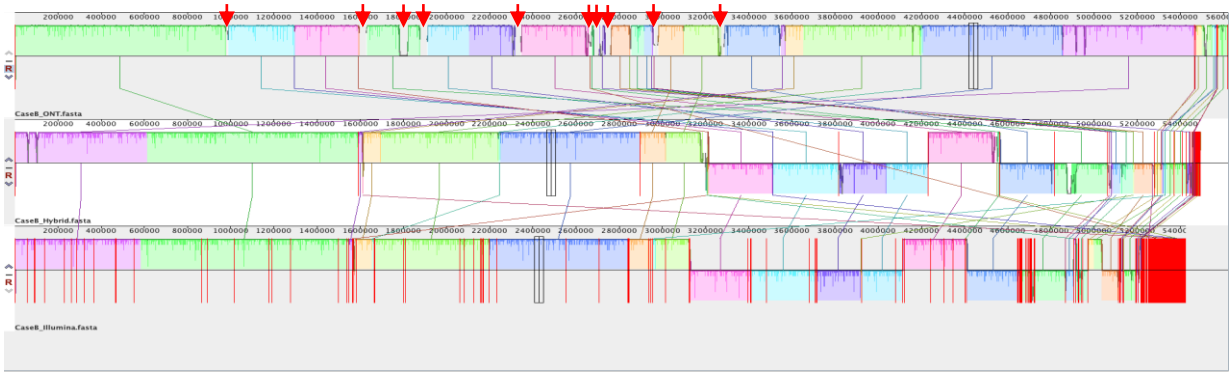
Case	# of contigs in ONT-only assembly (size bp)	# of contigs in hybrid assembly (size bp)	# of contigs in Illumina-only assembly (size bp)
A	5 (5,725,666 bp)	25 (5,506,670 bp)	668 (5,449,735 bp)
B	4 (5,620,611 bp)	34 (5,491,608 bp)	575 (5,424,436 bp)

Table 1 – Table showing the number of contigs generated and size of assembly for each assembly method for both cases.

Supplementary Figures



Supplementary figure 1 – Mauve alignment showing regions of similarity between the ONT-only, hybrid and Illumina-only assemblies (order descending) for Case A. Also showing the chromosomal regions in the ONT-only assembly that did not match the other assemblies (red arrows).



Supplementary figure 2 – Mauve alignment showing regions of similarity between the ONT-only, hybrid and Illumina-only assemblies (order descending) for Case B. Also showing the chromosomal regions in the ONT-only assembly that did not match the other assemblies (red arrows).

Reviewer reports:

Reviewer #1: This manuscript describes a comparative analysis of Illumina and Nanopore sequencing, evaluating their usefulness for phylogenetic analysis, and for identifying genetic variants in outbreak situations.

The outcome of the research is somewhat surprising, given the expectation that Illumina sequencing represents the current gold-standard in sequencing accuracy. When eliminating systematic variation in base sequences caused by methylation, Nanopore sequencing appears to have similar accuracy to Illumina sequencing for the purpose of variant categorisation. When methylation is considered as a important feature, Nanopore sequencing demonstrates both a greater detection ability, and a faster turnaround time compared to Illumina sequencing. I am pleased to note that the supporting data was available at the time I carried out my review, and also pleased to be given the opportunity to approve this manuscript for publication.

I was specifically asked by the editors to state whether this represents "the state-of-the-art in terms of what this platform can do." It would be underselling the impact of these results to say no, that the current basecalling technology is better than what is presented in this paper. Due to the rapid advancement of nanopore sequencing technology in hardware, software, and chemistry, the yield and quality of results obtained from nanopore sequencing will be better than what is in *any* publication, even at the time when a manuscript is submitted for review.

I recall seeing (and commenting) on David, Claire, Kathie, and Timothy's poster presented in April 2018, which seems to have been a similar (if not the same) study [https://twitter.com/gingerdavid92/status/987947325086666753]. This was the first study I'd seen that explicitly compared Illumina and Nanopore sequencing for phylogenetics [I accept there may be others that I haven't seen], and I'm pleased to see that they have incorporated an explicit analysis of methylation signals since then.

People have previously looked at phylogenetic trees for outbreak tracking with Nanopore sequencing (e.g. https://doi.org/10.2807%2F1560-7917.ES.2018.23.12.17-00140 [essentially cited in ref#10]), at accuracy estimates for Nanopore basecalling (e.g. https://doi.org/10.1101/543439), at hybrid isolate assembly from barcoded Nanopore and Illumina reads (e.g. https://doi.org/10.1099%2Fmgen.0.000132 [cited]), and at comparing clinical turnaround time for Nanopore vs Illumina (e.g. https://doi.org/10.1128/JCM.02483-16), but this paper puts it all together into something that is still of substantial interest to the research community, as demonstrated by the social media impact of their preprint (https://doi.org/10.1101/570192).

In short, this manuscript is an excellent demonstration of what nanopore sequencing is capable of, represents the state-of-the-art (as I understand it) for public health investigations as presented in published papers, and I look forward to seeing more studies like this in the future.

Additional comments / questions:

1. Results, Tables 1/2 line 194-200 - Could you please either add in the legend that these SNPs were homoplasmic (very unlikely for ONT, somewhat possible for Illumina), or add the depth of the reference SNP bases to the table?

We have added the depth of sequencing for the final discrepant SNPs in to Table 2.
We have identified what SNPs in the were homoplasmic (5/7 Illumina variants) and a line in the text.

2. Methods, line 348 - These were barcoded reads that were processed through Porechop, which I understand can identify and filter out chimeric reads. Do you know how many reads were chimeric (we've typically observed <0.5% chimeric reads from rapid adapter preps, about 4% from ligation preps)?

We have added a line in the methods with these figures.

3. Discussion, line 239 - It is interesting to see from Figure 1 that all the nanopore data analysis was completed before the sequencing run had ended. Maybe this could be emphasised here: "within 377 minutes (i.e. over 20 hours *before* the sequencing run was scheduled to finish)."

We have added a short statement emphasising the difference between technologies.

4. Discussion, lines 250-259 - The final sentence doesn't seem to match the general idea of this paragraph. The paragraph is about single-base accuracy for single molecules (note: Illumina never sequences a single molecule to generate a base call), whereas the last sentence is about phylogenetics. I'd be happier if this paragraph were deleted entirely, as phylogenetics and error are also discussed in the next paragraph.

We have removed the last sentence from this paragraph. We think it is important to keep the current limitations to show what was state of the art at the time of this publication

5. Discusion, line 283 - "long reads... workflow is" -> "longreads... workflow are"

This has now been corrected.

6. Discussion, line 303 - "up-dates" -> "updates"

This has now been corrected.

7. Figure 1 - Why were different methods used for DNA extraction (i.e. Promega Wizard vs Manual lysis / Qiagen Qiasymphony)?

The Qiasymphony utilises a magnetic beads in beating protocol that causes DNA fragmentation leading to a decrease in high molecular weight DNA molecules and thereby sub-optimal for long read sequencing. Therefore, a commercial gDNA extraction kit with modifications to attempt to keep the DNA integrity as high as possible and thus generate longer reads. We have added a sentence in the text to reflect the motivation of DNA extraction method.

8. Figure 2 - The numbers are difficult to read. Could the axis text be made larger?

We have enlarged the text for the axis for both graphs in Figure 2.

9. Figure 4 - This should be a line graph (similar to figure 3). The points represent sampling of potential cutoff scores along a continuous distribution, and the score represents a single value rather than count data.

We have modified the figure accordingly

10. Figure 5 - Table 1 (Line 165-166) mentions that the total number of discrepant variants for case A and B is 266 and 101 respectively. This doesn't match the percentages and totals represented in Figure 5. I would expect that the Total line for A/B in Figure 5 should be 97.7% and 93% respectively, indicating that transitions comprised that proportion of the total variants. It would be useful to refer back to Tables 1 & 2 in the text for the other discrepant variants.

We have clarified this in the legend in figure 5. In table 1 we are showing the total number of discrepant positions within both Illumina and ONT vcfs, however in figure 5 we are discussing only the variant positions in the ONT data alone that were determined to be methylated. Figure 5 should equate to the row titled '# of discrepant variants with methylated positions masked' in Table 1.

11. Figure 6 - What do the numbers represent? It is not clear from the figure legend. These are presumably not bootstrap values, as they have a consistent ordering from top to bottom.

These values represent each respective case's SNP differences between the Illumina and ONT as demonstrated on the tree. We have updated the figure (6) legend to now include this clarification.

General questions:

Given that the Nanopore technology has improved in a number of different areas since this investigation was carried out (e.g. 9.4.1 Series D Flow cells, Field sequencing kit and/or RBK004, flip-flop basecaller), what (if anything) would be done differently if you had the opportunity to do this again?

Of your suggestions, the only one likely to make a significant difference would be re-basecalling with the most up to date flip-flop basecaller, we would hypothesise that this would reduce the number of SNP differences if it accounts for methylation better than previous basecalling algorithms. Another advancement is the new R10 pores which aims to improve the consensus accuracy is about 99.999% which again would reduce the number of total SNP differences but will most likely not account for methylation (unless a trained basecaller is also developed with these pores to account for methylation).

Are the assemblies available? I can't see anything about the assemblies in the "Availability" section.

We have now uploaded our assemblies to NCBI and have added a comment (with accession numbers) in the "Availability of supporting data" section (lines 409-416).

Reviewer #2: In their paper, Greig et al. compare the performance of Oxford Nanopore Technology (ONT) and Illumina sequencing in identifying, subtyping and classifying clinical isolates in the context of outbreak investigation. This study is of considerable interest to both research and medical communities in two key aspects. Firstly, it provides an in-depth assessment of the performance of ONT versus the current sequencing standard Illumina Technology, and identifies the main mechanistic reason for discrepancies between the technologies (DNA methylation), which would be of value in optimizing and improving Nanopore analysis workflows. Secondly, they demonstrate that their real-time ONT analysis pipeline is able to rapidly provide diagnostic calls (species identification, serotyping etc) with comparable accuracy to Illumina sequencing in a fraction of the time, which has major potential applications in outbreak investigation.

Given the utility of this study, I would be happy to recommend it for publication - please consider the following points to possibly improve it further.

1) Are the methods appropriate to the aims of the study, are they well described, and are necessary controls included?

1. The sample size of 2 isolates is small, but justified given the context of the study (2 urgent cases of children with HUS admitted to the same hospital on the same night). However, if the authors have any other isolates sequenced (in particular, a reference isolate closely related to the reference genome) that allow for the comparison of Illumina/ONT, they could include it as supplemental information to improve the robustness of their assessment.

Currently we have only processed these two STEC O157:H7 samples using this methodology, we are hoping to follow up this manuscript in the future on a larger set of samples.

2. Run parameters for bioinformatics tools are well optimized and described. It would also be of major help to the community if the authors are willing to share the code for their real-time analysis pipeline.

Each component/tool was run individually during this study. We have not developed an automated pipeline though this is something we plan to develop in the future.

3. The authors adequately discuss the limitations of ONT relative to Illumina sequencing with respect to their application in rapid diagnosis.
   Fig 1/Methods - In the comparison of the ONT/Illumina workflows, we note that two different DNA

extraction methods are used (manual + QiaSymphony cleanup for Illumina, Promega Wizard Genomic DNA Purification for ONT). Are the methods interchangeable for the purposes of the workflow?

See point 7 for reviewer 1

2) Are the conclusions adequately supported by the data shown?
Analyses are generally robust and well-supported, but we would like clarification on the following points:

4. Line 214 - When comparing the case B ONT sequence with the 3 concurrent outbreak isolates, was it compared against Illumina sequences, or Nanopore sequences? If the comparison was between ONT and Illumina sequences, the discordance might arise from differences in the base-calling/software methods, and might disappear if all isolates were sequenced with ONT and compared directly (or would the high error rate preclude a valid comparison?) Please clarify and comment.

The outbreak case B sequenced via ONT was compared to Illumina sequenced isolates (and the equivalent case B sample sequenced via Illumina). We believe that the observed differences are inherent errors in both technologies. Our hypothesis is in agreement with yours, that comparing ONT to ONT sequenced outbreak strains would remove these discrepancies.

5. Line 216-218 - Given that 7 SNPs is not too dramatic a difference one could still make the case that the cases are quite plausibly linked. Would you be able to set an approximate SNV threshold for concluding genetic linkage?

Currently we use 5 SNPs as a proxy to infer genetic linkage or sharing the same epidemiological source with Illumina sequenced strains, through extensive validation from sequencing known outbreaks. With ONT sequencing and due to the lack of background sequenced samples we are unable to comment on an "appropriate" SNP threshold. We would have to take into account comparing ONT to Illumina data, Illumina to Illumina data and ONT to ONT data to decide if each type of comparison requires a different threshold or if we could set a general one to cover all comparisons.

3) Please indicate the quality of language in the manuscript. Does it require a heavy editing for language and clarity?

6. The language of the manuscript is quite clear. Please correct "manufactures instructions" to "manufacturer's instructions".

This has now been corrected at each use.

7. I also feel that the title of the manuscript downplays the speed and relative accuracy of the ONT diagnostic pipeline - the focus of the title should not be on the comparison of SNVs, but rather the comparison of the overall performance of the two methods. A title reflecting this and highlighting the rapid, real-time analysis capability of ONT-based diagnostics would be able to better capture reader interest and increase the impact of the manuscript.

8. Similarly, the abstract should be edited to emphasize the speed and real-time analysis capability.

We feel that the current title is appropriate for this study as the emphasis was that in conjunction with the nanopore being a rapid, real-time portable sequencer the current dogma is that variant calling is currently out of scope for this technology due to the high error rate. This manuscript refutes that dogma.

4) Are you able to assess all statistics in the manuscript, including the appropriateness of statistical tests used?
Yes - our group has experience analysing similar datasets. The precision/recall analysis performed is straightforward and appropriate.

(This manuscript was co-reviewed with Weizhen Xu, a postdoctoral fellow in my research group)

Reviewer #3: The authors examine the feasibility of using Nanopore sequencing to characterise single nucleotide variants in clinically relevant outbreaks. The authors present an interesting and relevant comparison of sequencing technologies given the rapid uptake of WGS as a clinical diagnostic tool. The data set is clearly described and accession code for all data submitted to the SRA are available in the manuscript and online. The methods are well described and all software/Bioinformatic tools are available online.

1. Although it can be addressed, my main concern with the manuscript is that the research was part funded by Oxford Nanopore. I believe the authors have not fully addressed the limitations of the Nanopore sequencing technology e.g. Cost, variability in throughput, etc. I? have highlighted some of these issues below. Platform limitations will also need to be included in the discussion

We have been more explicit to the funding received by Oxford Nanopore in the Acknowledgments. We have a paragraph on the limitations of ONT sequencing in terms of read accuracy and therefore variant detection which is the focus of this paper.

2. In the abstract and results the authors make a comparison of Illumina and ONT workflows of the time taken from DNA extraction to availability of results. The authors state that typing data (Shiga toxin subtyping and serotyping) was available within 7 hours while with Illumina it took ~40 hours to get these results. How do these time frames compare to standard laboratory based typing techniques that might be available in a diagnostic/pathology lab?

This paper covers the comparison between current methods deployed in the national reference laboratory – WGS by Illumina – with an alternative sequencing methodology ONT. It is out of scope of the paper to consider methods deployed in diagnostics e.g PCR

3. I would like to see a comparison of the sequencing costs for Illumina and Nanopore sequencing. A comparison against standard laboratory based typing techniques might also be beneficial to a broader audience (i leave that to the authors to decide).

Costing the sequencing technologies is not the focus of this paper and a would need to be a paper in its' own right considering labour / non-labour cost, deployment models etc. Also the value of such a comparison is incredibly time limited.

4. The authors state that the genetic relatedness of isolates could be determined at ~6 hours. I assume by genetic relatedness we are talking about variant/SNP typing. Can the authors explain how variants could be determined at 6 hours yet a MLST profile for Case B could not be determined until 10 hours had passed? Surely an inability to determine a MLST type indicates that the genome has not yet been sufficiently covered and it therefore unsuitable for variant typing.

Our SNP typing process requires an average genome coverage of 30x, the ONT sequencing took 6 hours to achieve this. As a result, as soon as 30x coverage is passed, this process can begin. Whereas, when sequencing the seven MLST genes, we are looking for enough coverage of those genes so that krocus can give us a confirmed result. This typically takes much longer, the last read was aligned to the seventh MLST gene at about 10 hours to then generate a full MLST result.

5. Additionally, in order to be cost effective multiple samples would need to sequenced on the same flowcell at the same time. Can the authors comment on what impacted multiplexing might have on the time frames described here?

This is correct, it is standard to multiplex several isolates per flow-cell. The higher the degree of multiplexing the increased pore competition we would expect the time to receive results per sample to increase. We have not performed this comparison. We have added this to the discussion

6. Can the authors comment as to why a number of regions (all describes as prophage with the exception of 1 in case A) were only present in the ONT-only chromosome assemblies? I find it odd that these regions were not present in the Illumina sequence data and are also absent from the

hybrid assembly. can the authors comment on why this might be the case and what impact that might have for genome sequencing and assembly strategies moving forward?

The main reason for the smaller assemblies in the Illumina and hybrid approaches is the large amount of paralogous sequences in STEC O157 encoded on cryptic phage. These sequences (which are longer then the Illumina read length) collapse into a single contig or are broken up into many small contigs with only a single copy when in reality they are multi-copy in the genome. This results in smaller genomes when Illumina reads alone or as a hybrid.

7. With regards to the genome assemblies of case A and Case B can the authors provide information on the number of erroneous indels that were present in the Nanopore assemblies? I assume these errors were polished out but did the authors only use Nanopore sequence data or was Illumina data also required.?

To keep the comparison as true as possible we only polished the ONT assembly with ONT data using Nanopolish. It is difficult in this case to quantify how many of the indels associated in the ONT assembly are correct or conversely incorrect in the Illumina assembly as many fall in prophage regions.

8. Line 129: Remove the MLST allele numbers

This has now been corrected.

9. Line 385: form -> for(?)

This has now been corrected.

10. Table 1 is very hard to interrupt. Consider restructuring the table.

We have reformatted the table to make the breakdown of SNPs clearer.