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 E Gharbia and Timothy J Dallman.

Gastrointestinal Bacteria Reference Unit, National Infection Service, Public Health England. London.

@gingerdavid92

Public Health England

Protecting and improving the nation's health

INTRODUCTION

- Shiga toxin-producing *Escherichia coli* (STEC) O157 is a zoonotic, foodborne pathogen defined by the presence of phage-encoded Shiga toxin genes (*stx*). Symptoms range from severe bloody diarrhoea to Haemolytic Uremic Syndrome (HUS). The STEC O157:H7 genome is typically 5.4 Mb in size, and a high proportion (>10%) of the genome comprises mobile elements, and prophage. The genome also typically including plasmids.
- During a foodborne outbreak investigation, there is often a need for a rapid typing result in order to confirm a microbiological link between an isolate in the implicated food and the outbreak strain isolated from clinical cases. Oxford Nanopore Technologies (ONT) offers a range of rapid sequencing platforms from the portable MinION to the more high throughput sequencers, GridION and PromethION.
- Currently, both short and long read technologies are being 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.
- We compared Illumina and ONT sequencing data from two isolates of STEC O157:H7 to determine whether the same single nucleotide variants were identified when mapped to references sequences in an established database.

METHODS

- Illumina sequencing:
  All DNA was extracted using the Qiasymphony platform (Qiagen). The sequencing library was prepared by fragmenting and tagging the purified gDNA using the Nextera XT DNA Sample Preparation Kits (Illumina) and sequenced on the Illumina HiSeq 2500 platform.
- Nanopore sequencing:
  All genomic DNA was extracted and purified using the Promega Wizard Genomic DNA Purification Kit with minor alterations. Library preparation was performed using the Rapid Barcoding Kit - SQK-RBK001 (ONT) with all three samples being barcoded and pooled (equimolar). The prepared library was loaded on a FLO-MIN106 R9.4 flow cell (ONT) and sequenced using the MinION for 24 hours.
- Bioinformatics:
  Nanopore data was basecalled using Albacore and Illumina and Nanopore data were processed through their respective pipelines shown in Figure 1.
- Nanopore pipeline optimisation:
  A precision/recall analysis was performed to determine which method of prophage masking should be used. Also a F1 score of the consensus threshold to call variants. Any methylation detection was performed using Nanopolish.

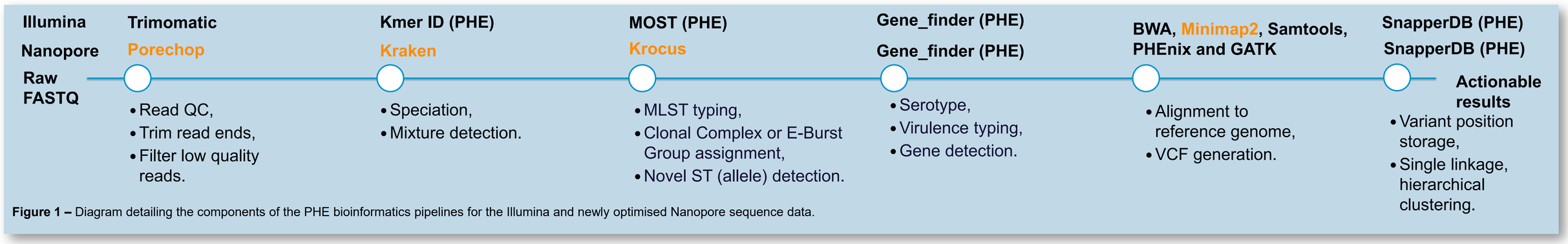
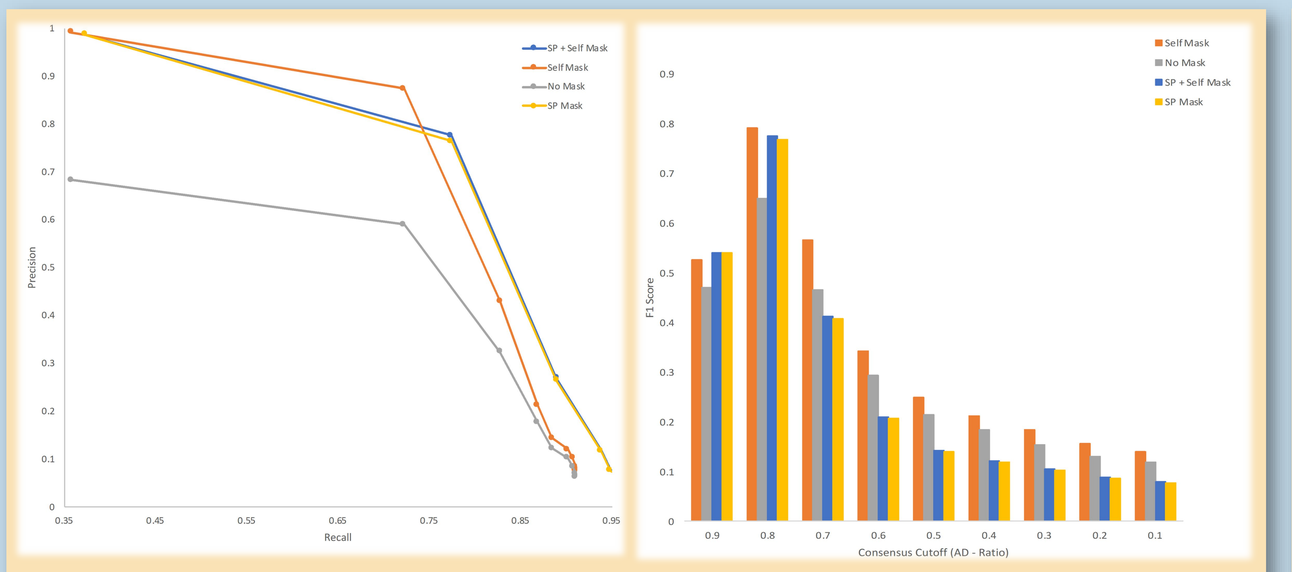


Figure 1 – Diagram detailing the components of the PHE bioinformatics pipelines for the Illumina and newly optimised Nanopore sequence data.

RESULTS



Figures 2 and 3. Left - Figure showing the precision/recall analysis for each methodology of prophage masking. Right – F1 score of each method of prophage masking in descending (by 10%) consensus thresholds.

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.	n/a	260	n/a	94
Final discrepant variants.	5	1	6	1

Table 1. Table showing the breakdown of the called variants and the number of variants masked for each case for each sequencing technology.

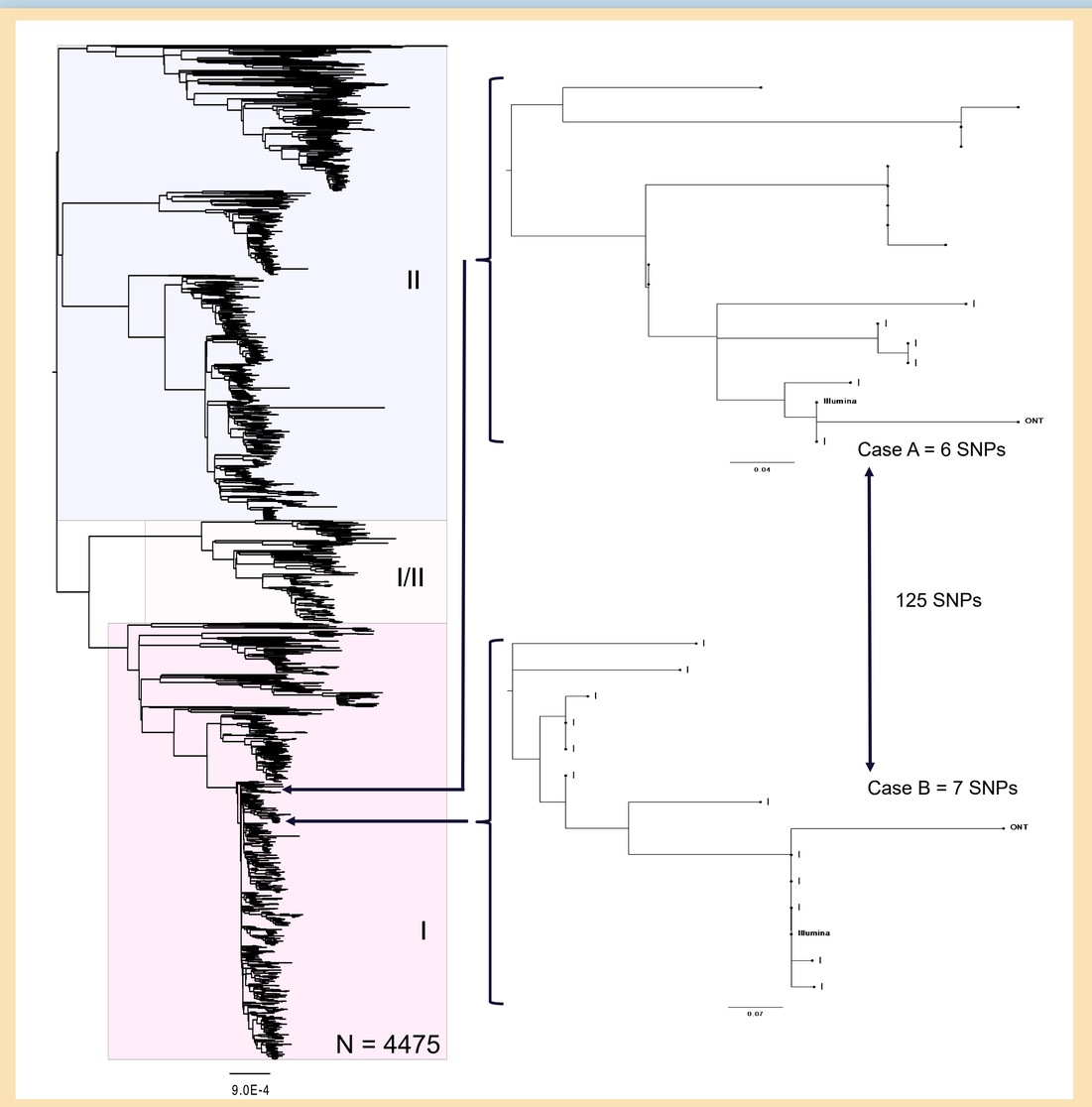


Figure 4 - Figure showing maximum likelihood tree of the STEC O157 population and relative positions of Case A and B for both Illumina and Nanopore technologies.

CONCLUSIONS

- We preliminary optimised our bioinformatics pipeline to handle Nanopore sequence data.
- We substituted Illumina specific bioinformatics components for Nanopore specific components.
- We optimised our variant calling parameters (Figures 2 and 3):
  - Consensus threshold (AD ratio) ≥ 0.8, depth ≥ 10 reads and mapping quality ≥ 30.
  - Prophage regions should be masked in the reference genome.
  - 5-methylcytosine methylation must be taken into account when calling variants in ONT sequence data.
- We demonstrate that you can rapidly discriminate between two samples for outbreak detection using Nanopore sequencing within a large database for Illumina sequences:
  - We demonstrate that Case A and B are on two difference clades and thus not related from the same source (not linked).
  - We show that Case B belongs to an outbreak.
- We show that you can use Nanopore sequence data for rapid SNP typing for outbreak detection despite the raw read and consensus error rates.

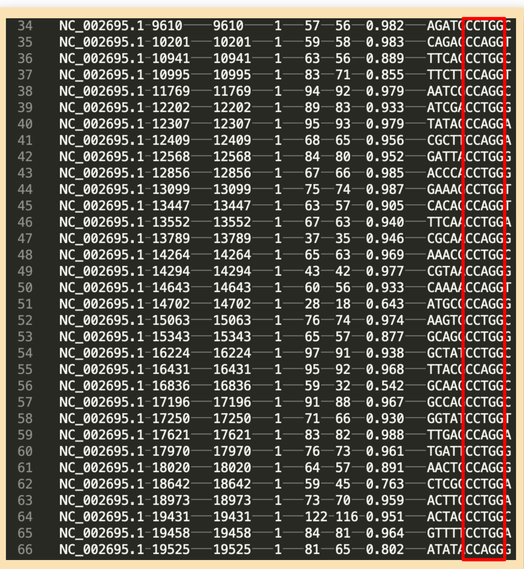


Figure 5 - Example of 5-methylcytosine methylation output from Nanopolish.

ACKNOWLEDGEMENTS

- I would like to thank Public Health England for performing the Illumina sequencing.
- I would like to thank the frontline NHS Laboratories for submitting the samples used in this study to GBRU at Public Health England.
- I would like to thank the NIRH HPRU GI for part-funding this project.
- I would like to thank Oxford Nanopore for funding my travel to ABPHM and for part-funding this project. Special thanks to Leila Luheshi and Divya Mirrington.

Bolger AM, Lohse M, & Usadel B. (2014). Trimmomatic: A flexible trimmer for Illumina Sequence Data. *Bioinformatics.* Btu170; Li H, and Durbin R. (2009) Fast and accurate short read alignment with Burrows-Wheeler Transform. *Bioinformatics.* 25:1754-60; Li H, *et al.* (2009) The Sequence alignment/map (SAM) format and SAMtools. *Bioinformatics*, 25, 2078-9; McKenna A. *et al.* (2010). The Genome Analysis Toolkit: a MapReduce frame- work for analyzing next-generation DNA sequencing data. *Genome Res.* 20, 1297–1303; Dallman T, *et al.* (2018) SnapperDB: A database solution for routine sequencing analysis of bacterial isolates. *Bioinformatics.* Bty212; Wick, R. Unpublished. https://github.com/rrwick/Porechop; Li, H. (2017). Minimap2: fast pairwise alignment for long nucleotide sequences. arXiv:1708.01492v5; 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; 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; 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.