Use of short and long read sequencing to investigate genetic relatedness during an outbreak of *E. coli* O157:H7.

David R Greig*[1,2,3], Timothy J Dallman[1,2,3], David L Gally[1,3], Saheer E Gharbia[1] & Claire Jenkins[1,2].

1) National Infection Service, Public Health England, London, UK. 2) NIHR HPRU in Gastrointestinal Infections. 3) Division of Infection and Immunity, The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, UK.

@gingerdavid92

Public Health England

Protecting and improving the nation's health

ROSLIN

INTRODUCTION

- Shiga toxin-producing *Escherichia coli* (STEC) are a group of zoonotic, foodborne pathogens defined by the presence of phage-encoded Shiga toxin genes (*stx*) [1]. STEC cause gastrointestinal disease in humans and symptoms include severe bloody diarrhoea, abdominal pain and nausea. In 5-15% of cases infection leads to Haemolytic Uremic Syndrome (HUS), characterised by kidney failure and/or cardiac and neurological complications [1].
- STEC O157:H7 genomes range from 5.4Mbp to 5.6Mbp in size, and a high proportion (9-15%) is comprised of mobile genetic elements and prophages [2].
- Due to the limitations of short read sequencing technologies in handling the homologous regions of the STEC chromosome, information and context regarding inter and intra variation in prophages, structural variation and context surrounding plasmid content is lost.
- We investigated an outbreak of nine cases of Shiga toxin-producing *Escherichia coli* (STEC) O157:H7 linked to participation in a mud-based obstacle race [3]. Three additional isolates from cases who could not be linked to the race fell within the same 5 single nucleotide polymorphism (SNP) cluster as the outbreak cases.

- We used a combination of short-read Illumina and long read Oxford Nanopore Technology (ONT) sequencing data to quantify genetic relatedness of 12 isolates and to look for micro-evolutionary events in the core and accessory genomes.

METHODS

DNA extraction was performed using a Qiagen Qiasymphony followed by library preparation using the Nextera XP kit followed by sequencing on the Illumina HiSeq 2500.

DNA extraction was also performed, using Revolugen's Fire Monkey kit followed by library preparation using SQK-RBK004 (Rapid) kit and sequencing on the Oxford Nanopore Technologies (ONT) MinION on a FLO-MIN106D flow cell.

Nanopore basecalling, read trimming and read filtering were performed using Guppy v3.2.4 FAST, Porechop v0.2.4[4] and Filtlong v2[5] respectively.

Nanopore reads where assembled using Flye v2.8[6] and the draft was corrected suing Nanopolish v0.11.3[6] (ONT reads), Pilon v1.22[7] (Illumina reads) and Racon v1.3.3[8] (Illumina reads).

Prophages sequences were collected manually from annotated finalised assemblies using Prokka v1.14.6[9] and compared in a pairwise format using Mash v2.2.2 [10].

Both Illumina and Nanopore datasets were processed using SnapperDB v0.2.6 to determine relatedness as described in Greig *et al* 2019[11].

RESULTS

- A comparison of variant calling and SNP typing of outbreak samples between short or long read sequencing data, placed 10/12 samples on the phylogeny within a single SNP of its pair. Two samples were located on a longer branch due to ambiguous aligning of short reads to the reference sequence, related to an insertion element. (Figure 1).
- All samples harboured a single *stx2a*-encoding prophage that was more similar in structure *stx2c*-encoding prophages and was located at a Shiga toxin encoding bacteriophage integration (SBI) site commonly associated with *stx2c*-econding prophages (*sbcB*) (Figures 2,3).
- All 12 samples contained the same number of prophages in total (n=17). However, there was evidence of micro-evolutionary events within the prophage content (Figures 3,4).
- All samples contained a 94kbp IncFIB plasmid. A further three samples (588888, 588889 and 591229) contained an extra 60kbp IncI2 plasmid of which one sample contained an additional 69kbp IncX4 plasmid and two samples contained another plasmid 41kbp of unknown Inc type.
- The two samples isolated from the previous year (2017) had a 1.44Mbp inversion between phages 4 and 12 relative to genomes from the outbreak in 2018. The same two genomes had a second event between phages 9 and 10 reverting a 0.45Mbp portion of the genome sequence back into the same orientation as the remaining 10 genomes (Figure 3).

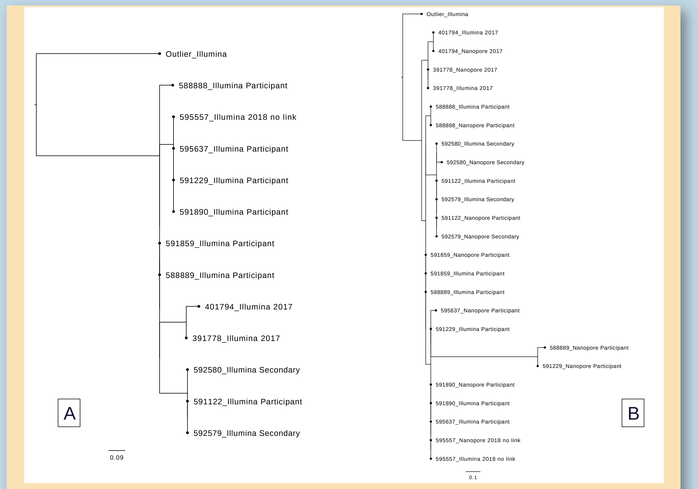

Figure 1. Maximum-likelihood phylogeny showing the outbreak cluster CC11 sub-lineage IIb (A). A second maximum-likelihood phylogeny showing both Illumina derived and Nanopore derived SNP-typing results for each of the outbreak samples (B).

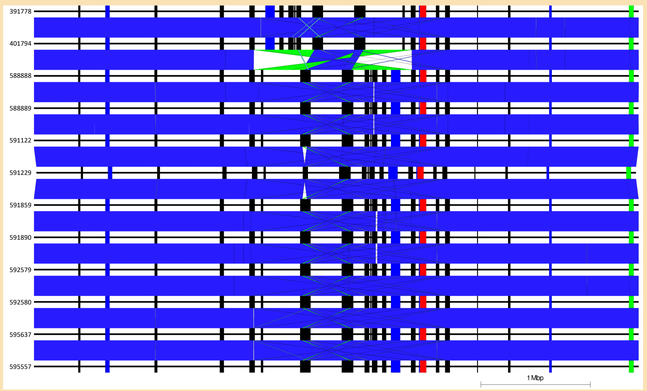

Figure 3. Easyfig[12] alignment showing the chromosome and loci of prophages in all samples in the outbreak in question. *Stx*-encoding prophage, Red; Prophage-like region, Blue; Locus of Enterocyte Effacement (LEE), Green and other non-*stx*-encoding prophages, Black.

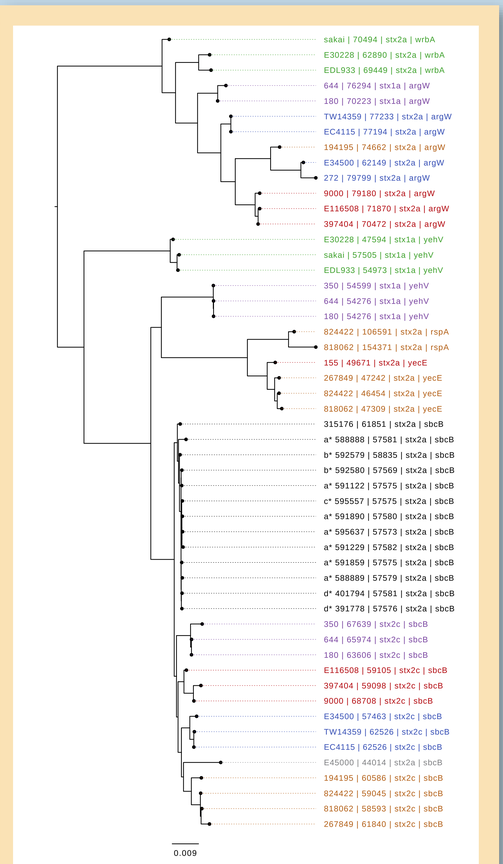

Figure 2. Neighbour joining tree based on Jaccard distances of *stx*-encoding prophages of publicly available samples and the outbreak samples sequenced in this study (labelled with * preceding strain ID).

Prophages are coloured by CC11 sub-lineage. Sub-lineage Ia, Green; Ib, Yellow; Ic, Red; I/IIa, Blue; I/IIb, Grey; IIa, Orange; IIb, Black and IIc, Purple.

A, race participant; B, secondary/household; C. geographically linked 2018; D, geographically linked 2017.

DISCUSSION and CONCLUSIONS

- Comparing Illumina with Nanopore sequencing data highlighted the difficulty of SNP detection associated with attempting to integrate both technologies. However, the analysis of the Nanopore data confirmed the close genetic relatedness demonstrated by the Illumina data. It also the highlights the need for correct masking of mobile genetic elements and methylated sequences relative to the reference genome used for variant calling.
- Despite all 12 samples being within the same 5-SNP single linkage cluster using Illumina sequencing data, analysis of the Nanopore sequencing data revealed variation in both plasmid and prophage content of the genomes.
- The *stx2a*-encoding prophage harboured by the outbreak strains was located at the loci associated with *stx2c*-encoding prophages, and this may have an impact on pathogenicity of the strain.
- With the development of long-read sequencing technology, we can now detect and describe both large- and small-scale structural variation and explore the effect on phenotype.
- The ability to characterise the accessory genome in this format is the first step to understanding the significance of these micro-evolutionary events and their impact on the evolutionary history, virulence, and potentially the likely source and transmission of this zoonotic, foodborne pathogen.

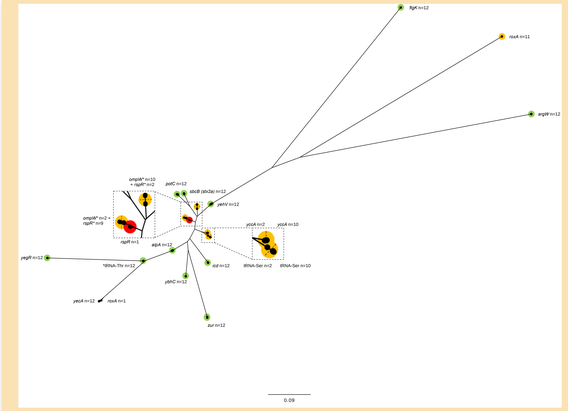

Figure 4. Neighbour joining tree based on Jaccard distances of all prophages within samples in the outbreak in question. Prophage clusters are coloured as follows: Green, shared between all samples (n=12); Yellow, shared between two samples or more and Red, unique to a single sample.

Clusters are labelled with the SBI of that prophage and the number of samples that contained that phage. * denotes compounded prophages.

ACKNOWLEDGEMENTS

The research was funded by the National Institute for Health Research Protection Research Unit (NIHR HPRU) in Gastrointestinal Infections at University of Liverpool in partnership with Public Health England (PHE), in collaboration with University of Warwick. The views expressed are those of the author(s) and not necessarily the NIHR, the Department of Health and Social Care or Public Health England.

NIHR | Health Protection Research Unit in Gastrointestinal Infections at University of Liverpool

REFERENCES

1) 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;143:3475-87. doi: 10.1017/S0950268815000746. **2)** Croxen MA, Law RJ, Scholz R, Keeney KM, Wlodarska M, Finlay BB. Recent advances in understanding enteric pathogenic *Escherichia coli*. Clin Microbiol Rev. 2013;26:822-80. doi: 10.1128/CMR.00022-13 **3)** Sharp A, Smout E, Byrne L, Greenwood R, Abdoollah R, Hutchinson C *et al*. An outbreak of Shiga toxin-producing *Escherichia coli* O157:H7 linked to a mud-based obstacle course, England, August 2018. Zoonoses Public Health. 2020;67(5):467-473. doi: 10.1111/zph.12744. Epub 2020 Jun 21. **4)** Wick R. Unpublished. https://github.com/rrwick/FiltLong. **5)** Wick R. Unpublished. https://github.com/rrwick/Porechop. **6)** 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. **7)** 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. **8)** Vaser R, Sović I, Nagarajan N, Šikić M. 2017. Fast and accurate de novo genome assembly from long uncorrected reads. Genome Res. 27(5):737-46. doi: 10.1101/gr.214270.116. **9)** Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics 2014;30:2068–2069. doi: 10.1093/bioinformatics/btu153. **10)** Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH et al. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol 2016;17:132. doi: 10.1186/s13059-016-0997-x. **11)** Greig DR, Jenkins C, Gharbia S, Dallman TJ. Comparison of single nucleotide variants identified by Illumina and Oxford Nanopore technologies in the context of a protentional outbreak of Shiga toxin-producing *Escherichia coli*. Gigascience. 2019;8(8): https://doi.org/10.1093/gigascience/giz104.