Nanopore sequencing reveals the hidden intra-outbreak accessory genome variation of Shiga toxin-producing *E. coli* (STEC) O157:H7.

UK Health Security Agency

David R Greig [*1,2,3], David L Gally[1,3], Saheer E Gharbia[1,4], Timothy J Dallman[5] & Claire Jenkins[1,2]

@gingerdavid92

1) UKHSA, London, UK. 2) NIHR HPRU for GI pathogens, Liverpool, UK. 3) The Roslin Institute, University of Edinburgh, Edinburgh, UK.
4) NIHRI HPRU for GED, Warwick, UK. 5) Utrecht University, Utrecht, Netherlands.

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 retrospectively investigated five outbreaks of Shiga toxin-producing *Escherichia coli* (STEC) O157:H7:

- Associated with consumption of contaminated leafy greens (n=18),
- Associated with consumption of contaminated mince beef (n=17),
- Associated with participation in a mud based obstacle course (n=12),
- Associated with attendance of a lambing event (n=10)
- Associated with consumption of raw drinking milk (n=23).

The ability to scrutinise the accessory genomes of pathogens provides insight to the dynamic nature of the accessory genome, acquisition and loss of virulence genes and antibiotic resistance determinants, the genomic context of mobile genetic elements and large chromosomal rearrangements, that may have public health implications.

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 v3-5 FAST, Porechop v0.2.4[3] and Filtlong v2[4] respectively.
- Nanopore reads where assembled using Flye v2.8[5] 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 v0.2.8 to determine relatedness as described in Greig *et al* 2019[11].

RESULTS

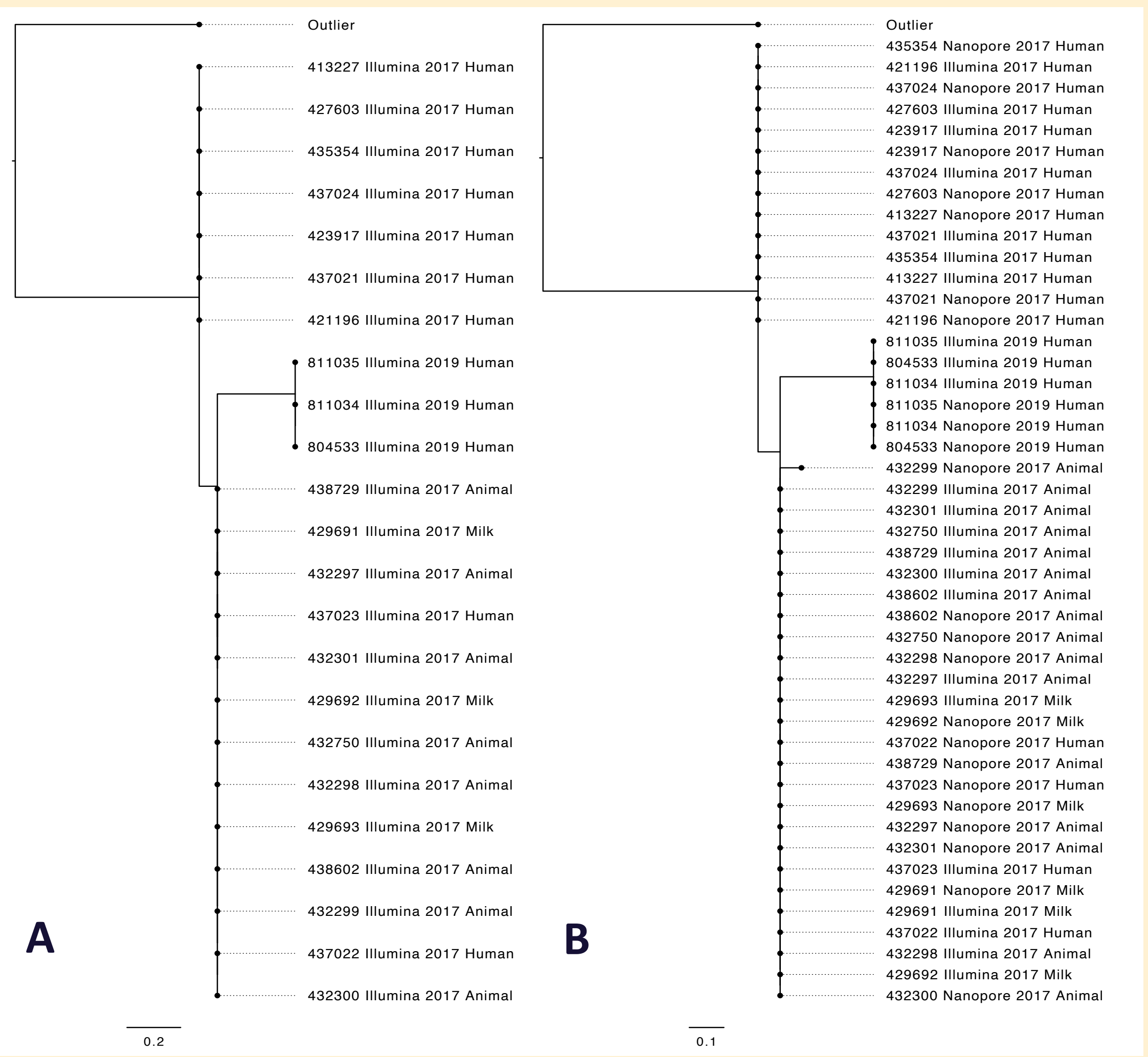

Figure 1. Maximum-likelihood phylogeny showing the raw drinking milk outbreak cluster (A). A second maximum-likelihood phylogeny showing both Illumina derived and Nanopore derived SNP-typing results for each of the outbreak samples (B).

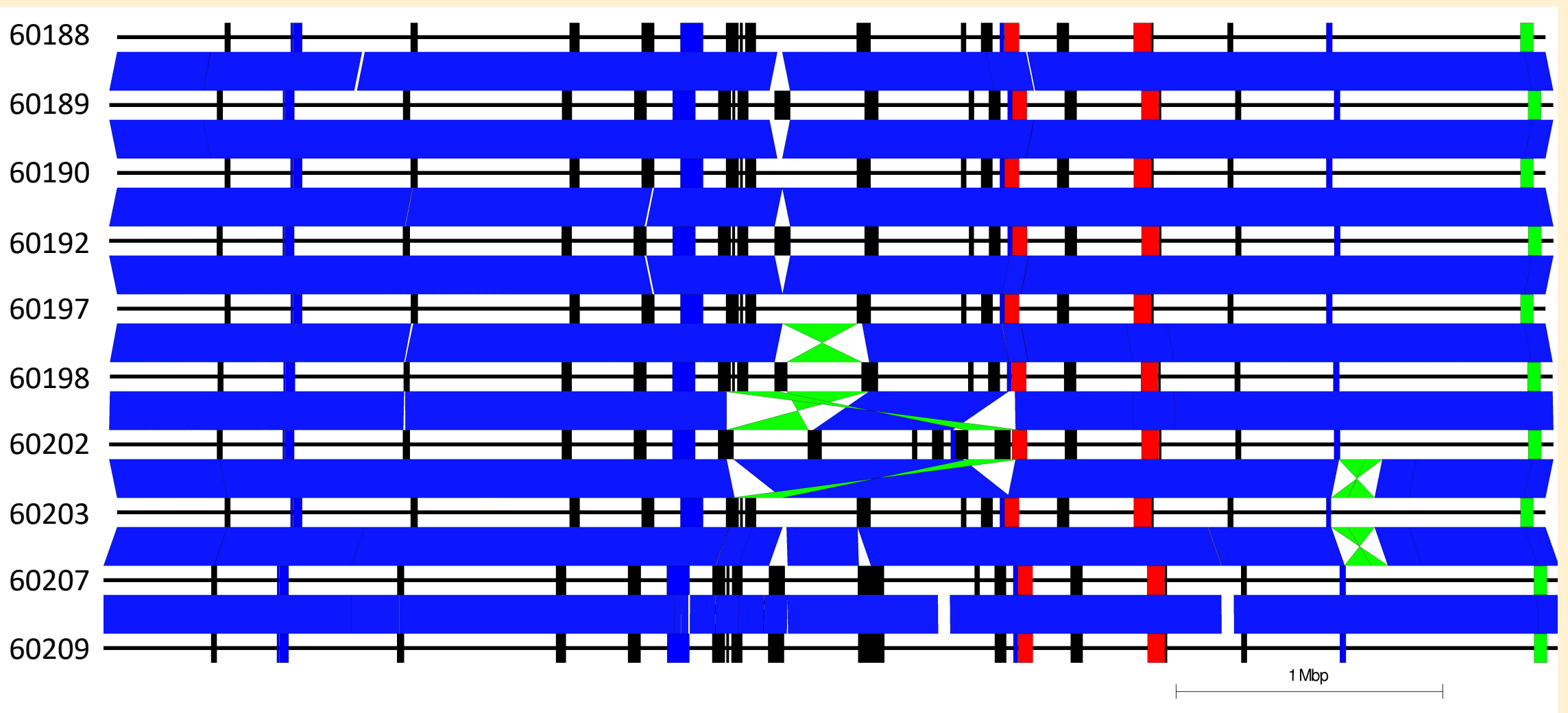

Figure 4. Easyfig[12] alignment showing the chromosome and loci of prophages in all samples in the petting farm associated outbreak. *Stx*-encoding prophage, Red; Prophage-like region, Blue; Locus of Enterocyte Effacement (LEE), Green and other non-*stx*-encoding prophages, Black.

- A comparison of variant calling and SNP typing of raw drinking milk outbreak samples between short or long read sequencing data, placed 23/23 samples on the phylogeny within a single SNP of its pair. Only one sample was a single SNP from its Illumina equivalent. (Figure 1).
- Nanopore sequencing enabled the comparison of *stx*-encoding prophages across outbreaks. This comparison showed that most *stx*-encoding prophages cluster based on the STEC CC11 sub-lineage of their host and *stx*-encoding bacteriophage integration site (SBI) (Figure 2).

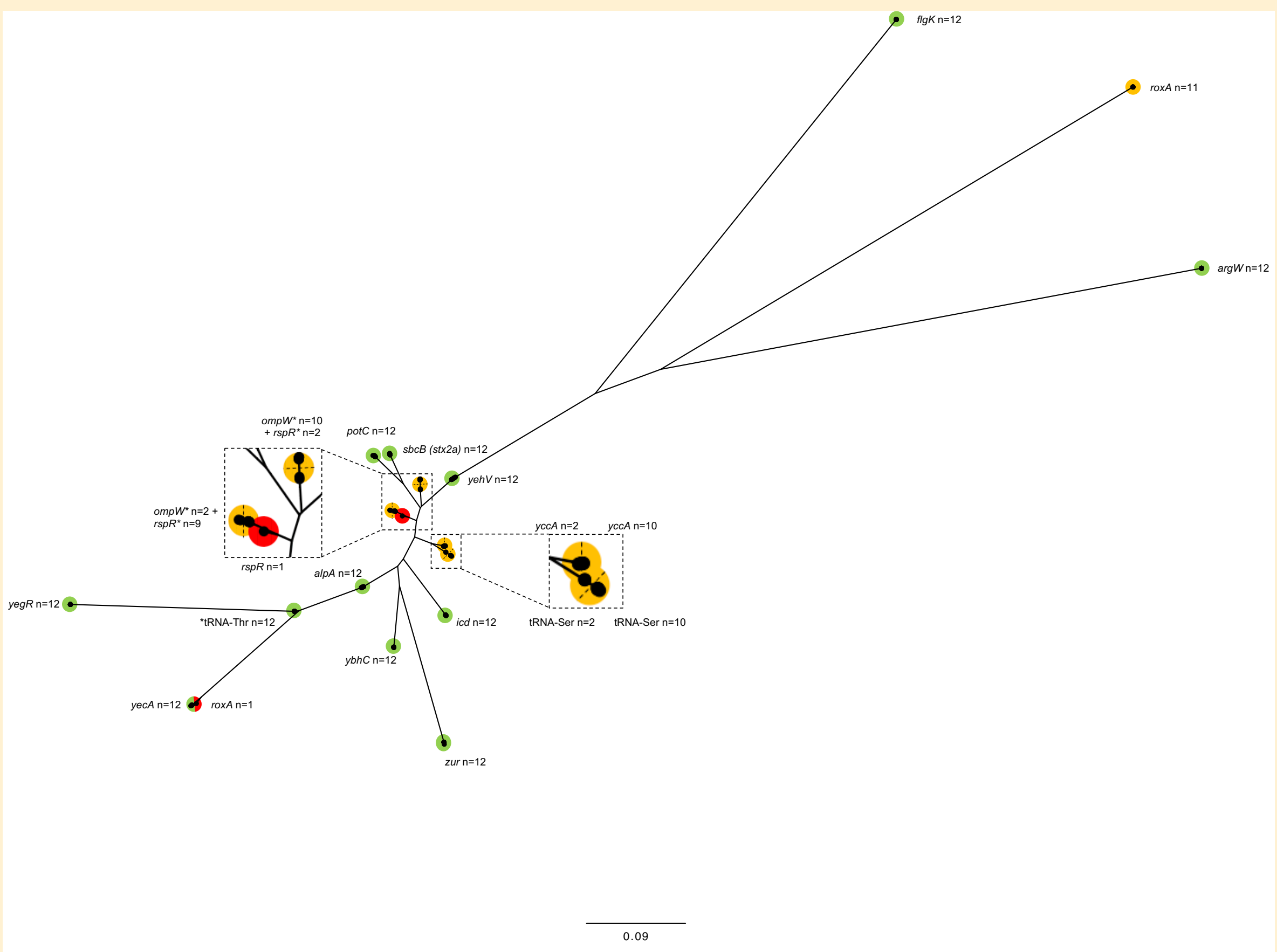

Figure 3. Neighbour joining tree based on Jaccard distances of all prophages within samples in the obstacle course associated outbreak. Prophage clusters are coloured as follows: Green, shared between all samples (n=23); Yellow, shared between two samples or more and Red, unique to a single sample.

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

- The prophage content of each outbreak including non-*stx*-encoding prophages was also variable. Food associated outbreaks showed a more conserved prophage content with animal contact and environmental (mud obstacle course) associated outbreaks displaying more prophage content variation. (Figure 3).
- Each outbreak had varying levels of micro-evolutionary events with some chromosome's being quite conserved and others containing many large chromosomal re-arrangements and translocations as in the petting farm outbreak (Figure 4).

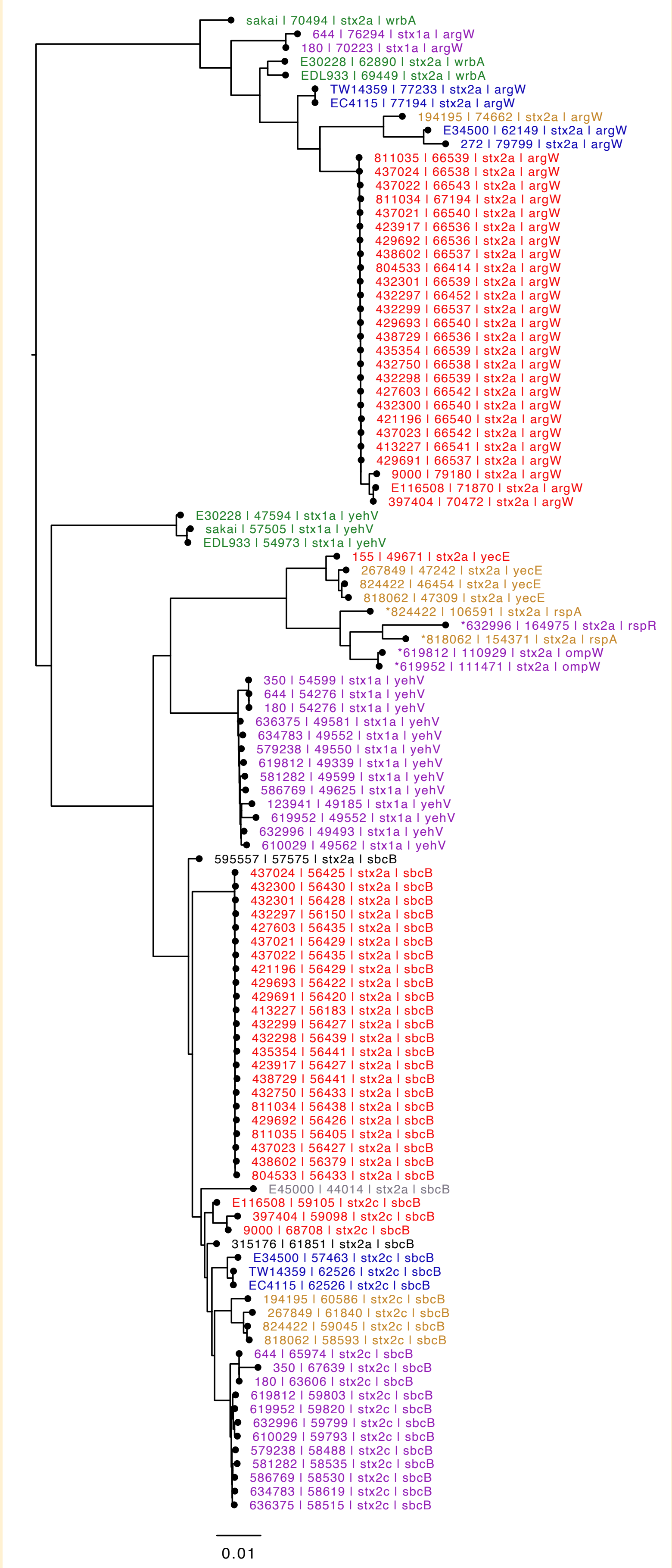

Figure23. Neighbour joining tree based on Jaccard distances of publicly available and raw drinking milk associated *stx*-encoding prophages.
Prophages are coloured by CC11 sub-lineage. Sub-lineage Ia, Green; Ib, Yellow; Ic, Red; I/IIa, Blue; I/IIb, Grey; IIa, Orange; IIb, Black and IIc, Purple.

DISCUSSION & CONCLUSIONS

- Nanopore sequencing can generate information in real time leading to faster generating of results and could help to implement public health actions faster.
- Nanopore sequencing can open the accessory genome of GI pathogens which is currently much more difficult with short-read sequence technologies.
- This ability will allow us to determine more information from the accessory genomes of GI pathogens including:
  - Detection and characterisation of prophage content.
  - Isolation and typing of plasmid content.
  - Detection of large-scale chromosomal rearrangements and other structural variation.
- The genomes of emerging highly-pathogenic strains can be characterised rapidly and aid in our understand as to why they are more pathogenic or emerging more successfully.
- 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.

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 UK Health Security Agency (UKHSA) formally Public Health England (PHE), in collaboration with University of Warwick. The views expressed are those of the author(s) and not necessarily those of the NIHR, the Department of Health and Social Care or UKHSA.

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

ROSLIN

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)** Wick R. Unpublished. https://github.com/rrwick/FiltLong. **4)** Wick R. Unpublished. https://github.com/rrwick/Porechop. **5)** Kolmogorov M, Yuan J, Lin Y and Pevzner PA. 2019. Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol. 37(5):540-546. doi: 10.1038/s41587-019-0072-8. **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. **12)** Sullivan MJ, Petty NK, Beatson SA. Easyfig: a genome comparison visualizer. Bioinformatics 2011;27:1009–1010. doi: 10.1093/bioinformatics/btr039.