What the phage? A genomic map of the prophage content of domestic UK Shiga-toxin producing *E. coli* (STEC) O157:H7 genomes.

UK Health Security Agency

David R Greig[1,2,3*] & Claire Jenkins[1,2]

@gingerdavid92

1) National Infection Service, United Kingdom Health Security Agency, London, NW9 5EQ, UK. 2) NIRH Health Protection Research Unit for Gastrointestinal Pathogens, Liverpool, UK. 3) Division of Infection and Immunity, The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, UK.

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-outbreak variation in prophages, structural variation and context surrounding plasmid content is lost. As a result, the prophage population of STEC O157 is largely unexplored.

In this project we used a combination of short-read Illumina and long-read Oxford Nanopore Technology (ONT) sequencing data to generate complete genome assemblies giving us access the accessory genome to characterise and compare the prophage content within domestic UK STEC O157:H7 .

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 v3.2.10 FAST – Guppy v6.3.8 FAST, Porechop v0.2.4[4] and Filtlong v2[5] respectively.

Nanopore reads where assembled using Flye v2.9[6] and the draft was corrected using four iterations of Racon v1.4.20[7] (ONT reads), Medaka v0.1.0[8] with an STEC-specific model (ONT reads), Pilon v1.23[9] (Illumina reads) and finally Racon v1.4.20[7] (Illumina reads).

Prophages sequences were collected manually from annotated finalised assemblies using Prokka v1.14.6[10] and compared in a pairwise format using Mash v2.2.2 [11].

RESULTS

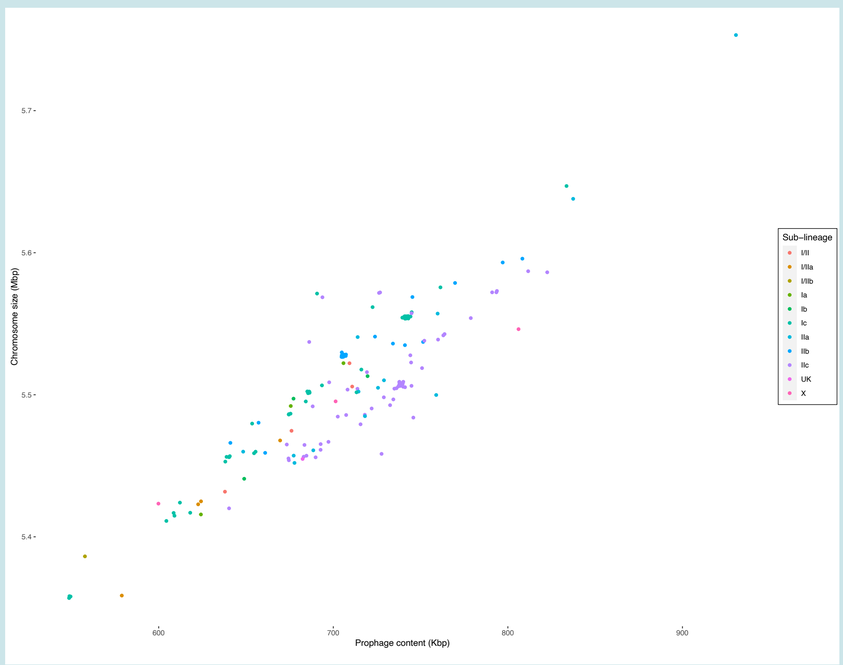

Figure 1.
Left – Chromosome size vs number of prophages.

Right – Chromosome size vs total prophage content (kbp).

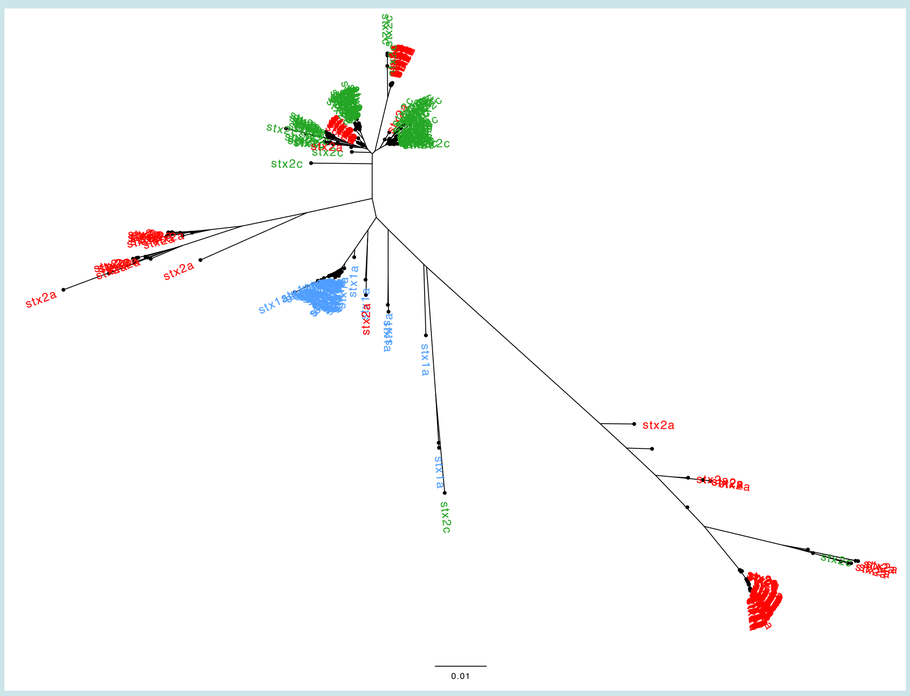

Figure 2. Neighbour joining tree based on Jaccard distances of all *stx*-encoding prophages (n=308) within the 176 STEC genomes sequenced. *stx2a,* stx2c and *stx1a* are coloured red, green and blue respectively.

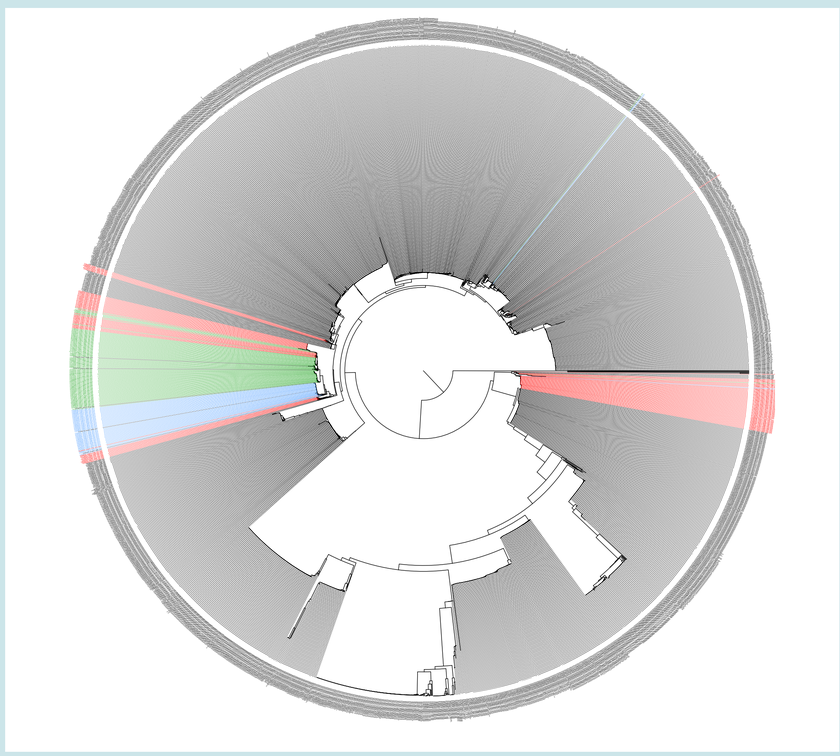

Figure 3. Neighbour joining tree based on Jaccard distances of all prophages (n=2,981) within the 176 STEC genomes sequenced.

- Long-read sequencing and bioinformatics processing produced 176 complete assemblies of STEC O157:H7. These genomes yielded 2,981 prophages in total of which 308 are *stx*-encoding prophages.
- Prophage content varied from 14 to 20 prophages per chromosome, with STEC O157:H7 belonging to lineage II having more prophages on average than those in lineage I and I/II (Figure 1).
- Prophages harbouring *stx1a* had a preferential *stx*-encoding bacteriophage integration site (SBI) of *yehV* (96.6%, 58/60) whereas *stx2c* had a preferential site of *sbcB* (98.25 112/114) and in most cases *stx2c* and *stx1a*-encoding showing little genetic diversity (Figure 2).
- *stx2a*-encoding prophages had the most diverse SBI sites, including *argW* (49.2% 66/134), *sbcB* (31.3% 42/134), *yecE* (8.2% 11/134) and this was reflected in terms of genetic distance as *stx2a*-encoding prophages showed much more diversity (Figure 2).
- There were 15 genomes that had multiple copies of *stx2a*-encoding prophages, inserted at different SBIs which would not have been detected via short-read sequencing.

DISCUSSION & CONCLUSIONS

Long-read sequencing allows for the characterisation of the accessory genome of STEC and thus allows us to detect and extract complete prophage sequences.

By cataloging the prophage content of STEC O157:H7 into a database is the first step into developing novel *in silico* methods of typing such as phage typing or determining if a given sample is domestic or non-domestic.

Accurately mapping the acquisition and loss of *stx*-encoding bacteriophages enables us to predict the virulence potential of the different STEC O157:H7 lineages, and to monitor for emerging threats to public health.

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 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)** Treacy J, Jenkins C, Paranthaman K, Jorgensen F, Mueller-Doblies D, Anjum M et al. Outbreak of Shiga toxin-producing Escherichia coli O157:H7 linked to raw drinking milk resolved by rapid application of advanced pathogen characterisation methods, England, August to October 2017. Euro Surveill. 2019;24(16):1800191. doi: 10.2807/1560-7917.ES.2019.24.16.1800191. **4)** Wick R. Unpublished. https://github.com/rrwick/FiltLong. **5)** Wick R. Unpublished. https://github.com/rrwick/Porechop. **6)** Kolmogorov, M. Yuan, J. Lin, Y. & Pevzner, P.A. Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol 37:540–546. (2019) doi: 10.1038/s41587-019-0072-8. **7)** 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. **8)** https://github.com/nanoporetech/medaka **9)** 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. **10)** Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics 2014;30:2068–2069. doi: 10.1093/bioinformatics/btu153.. **11)** 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.