Cóimeáil: A bioinformatics pipeline for the typing and characterisation of Shiga-toxin producing *E. coli* genomes sequenced via Oxford Nanopore Technologies.

UK Health Security Agency

David R Greig[1,2,3]*, Ella V Rodwell[1,2], Anaïs Painset[1] & Claire Jenkins[1,2]

@gingerdavid92

1) National Infection Service, United Kingdom Health Security Agency, London, NW9 5EQ, UK. 2) NIHR 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 variation in prophages, structural variation and context surrounding plasmid content is lost.

With the advent and development of long-read sequencing technologies, we can now generate single contiguous *de novo* assemblies of complex bacterial genomes containing homologous sequences. This facilitates the characterisation and typing of elements of the accessory genome of gastrointestinal pathogens, including those with a high bacteriophage content.

Here we present Cóimeáil, a python pipeline which utilises both the long-read sequences and complete contiguous assemblies to derive pathogen typing data in a dual format. The data output includes all the components derived from the analysis of short read data, specifically bacterial identification to the species level, multi-locus sequence typing (MLST), virulence and antimicrobial gene detection and assessment of strain relatedness.

However, by utilising the nature of long-read sequencing, Cóimeáil also delivers copy number detection of virulence and antimicrobial resistance (AMR) genes, prophage characterisation and plasmid typing.

METHODS / WORKFLOW

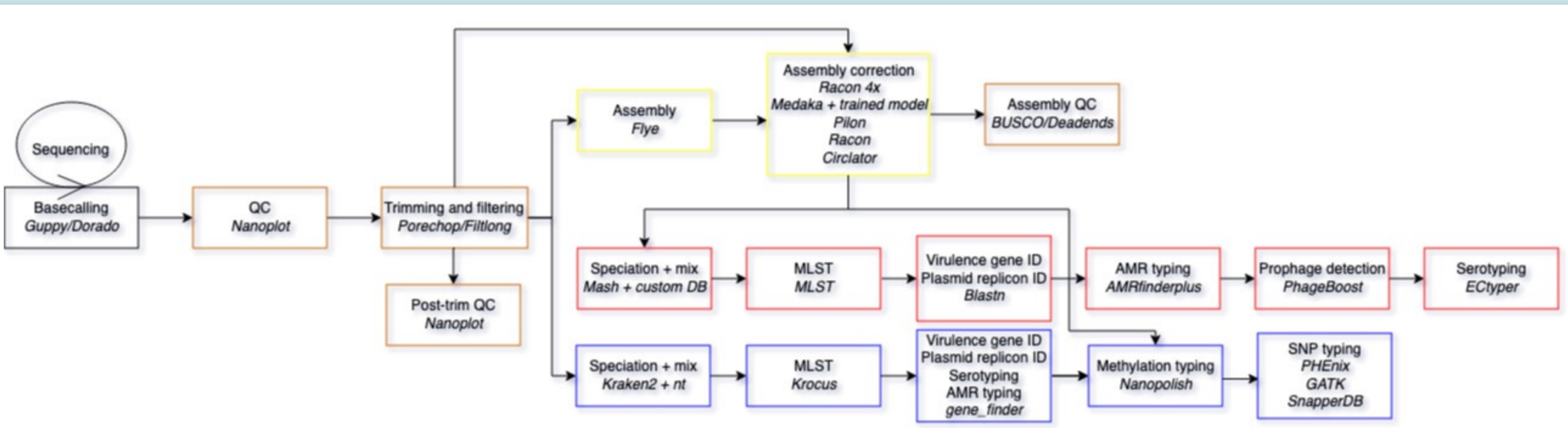

Figure 1. Basic Cóimeáil workflow showing data progression and types of results derived. Orange = Quality control steps; Yellow = Assembly; Red = Assembly-based typing; Blue = Read-based typing.

The typing results produced by Cóimeáil:

- Speciation – what species is this sample? [reads and assembly]
- Mixture detection [reads and assembly]
- Multi-locus sequence typing (MLST) typing [reads and assembly]
- Serotyping (somatic and flagellar antigen detection) [reads = presence/absence and assembly with dual detection]
- Virulence gene detection [reads = presence/absence and assembly provides with locus]
- AMR gene detection [reads = presence/absence and assembly provides with locus]
- Plasmid replicon detection [reads = presence/absence and assembly provides with locus]
- *Stx* subtyping [reads = presence/absence and assembly provides with locus + copy number]
- Prophage detection [assembly only]
- Methylation distribution on reads vs assembly of self.
- Structural Variant (SV) typing using reads vs assembly of self.
- SNP typing [reads only].

Run ID	Sample ID	Assembly-based typing results											Read-based typing results							
		Speciation	Mixture	*E. coli* Phylogroup	Serotype	Serotype warning	Blast-based Serotype	ST	MLST Profile	Plasmid replicon	*Stx* subtype	Virulence (*eae*)	Prophages	Speciation	Mixture	ST	MLST Profile	Serotype	Plasmid replicon	*Stx* subtype	Virulence (*eae*)
<Run ID>	<Sample ID>	*Escherichia coli*	N	B1	O45:H2	N	O45:H2	20	6,4,3,18,7,7,6	IncFIB (185kbp) IncFIC (107kbp)	*stx2f stx2f*	+	26	*Escherichia coli*	N	20	6,4,3,18,7,7,6	O45:H2	IncFIB IncFIC	*stx2f*	+

Table 1. An example of the results generated by Cóimeáil for a single STEC genome.

RESULTS

An initial validation set of 64 *E. coli* were processed through Cóimeáil and compared to current WGS (Illumina-based) and PCR typing methods at GBRU, UKHSA.

For assembly-based typing speciation matched 100%, 100% in sequence type assignment, 94% (60/64) in *E. coli* serotype determination when compared to current WGS methods.

Traditional WGS methods 89% (57/64) were outperformed by Cóimeáil in detection of the *eae* gene, a prominent virulence factor of STEC, by matching 100% to known PCR results of this gene for the validation samples.

Detection and differentiation of *stx* subtype matched 100% between Cóimeáil and current WGS methods in terms of gene presence and absence.

Cóimeáil also delivered additional typing results in terms of copy number gene detection for example in three samples multi-copy *stx* genes were detected (e.g. *stx2f/stx2f*). This is not possible with current methods as they are alignment-based only.

Read-based typing matched 100% speciation and 100% in sequence type assignment compared to current WGS methods. Other read-based components are still in development at time of writing.

Cóimeáil was designed to be modular and lightweight so the entire pipeline can be run on a standard 8GB RAM/4 CPU laptop in a sequential format.

DISCUSSION & CONCLUSIONS

- Cóimeáil operates by deriving typing data from both Nanopore reads and a long-read assembly. This provides the user with a mirror results set which can provide additional context where one dataset might struggle to derive a result alone. Cóimeáil, also utilises the long-read nature of the results to derive results that are not possible with short-read sequencing, including detection of copy number of important genes, gene localisation, detection of structural variation, detection of prophages and allows for downstream whole-genome/chromosome/plasmids comparisons.
- With the fast-moving field of long-read genomics it is difficult to accreditate a bioinformatics pipeline to a standard. Cóimeáil provides a framework for developing a locked-down version-controlled workflow. Additionally, Cóimeáil's only requirements are the Nanopore FAST5 and raw FASTQ files. This means that any previous Nanopore sequencing run can be re-processed at a later date with a single command.
- The ability to characterise the STEC accessory genome in this standardised format is the first step to understanding the significance of the newly derived accessory genome micro-evolutionary events and their impact on the evolutionary history, virulence, and potentially the likely source and transmission of this zoonotic, foodborne pathogen.
- Due to the modular nature of Cóimeáil, development of new components or an entire workflow for another gastrointestinal pathogen is possible. Currently we are developing the components/workflow to characterise and type *Shigella* genomes, with the long-term goal to be able to characterise all gastrointestinal pathogens that are processed by GBRU, UKHSA.

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) doi: 10.1017/S0950268815000746. **2)** doi: 10.1128/CMR. **3)** Wick R. Unpublished. https://github.com/rrwick/FiltLong. **4)** Wick R. Unpublished. https://github.com/rrwick/Porechop. **5)** doi: 10.1093/bioinformatics/bty149. **6)** doi: 10.1038/s41587-019-0072-8. **7)** doi: 10.1371/journal.pone.0112963. **8)** doi: 10.1101/gr.214270.116. **9)** doi: 10.7717/peerj.5233. **10)** doi: 10.1186/s13059-016-0997-x. **11)** Wick R. Unpublished. https://github.com/rrwick/GFA-dead-end-counter. **12)** Seemann T Unpublished https://github.com/tseemann/mlst. **13)** doi: 10.1038/nmeth.3444 **14)** doi: 10.1093/bioinformatics/bty212. **15)** doi: 10.1186/1471-2105-10-421. **16)** doi: 10.1038/s41598-021-91456-0. **17)** doi: 10.1093/nargab/lqaa109. **18)** doi: 10.1099/mgen.0.000728.