Cóimeáil: A Nanopore-based bioinformatics pipeline for the rapid typing and characterisation of gastrointestinal pathogens.

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

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.

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.

WORKFLOW

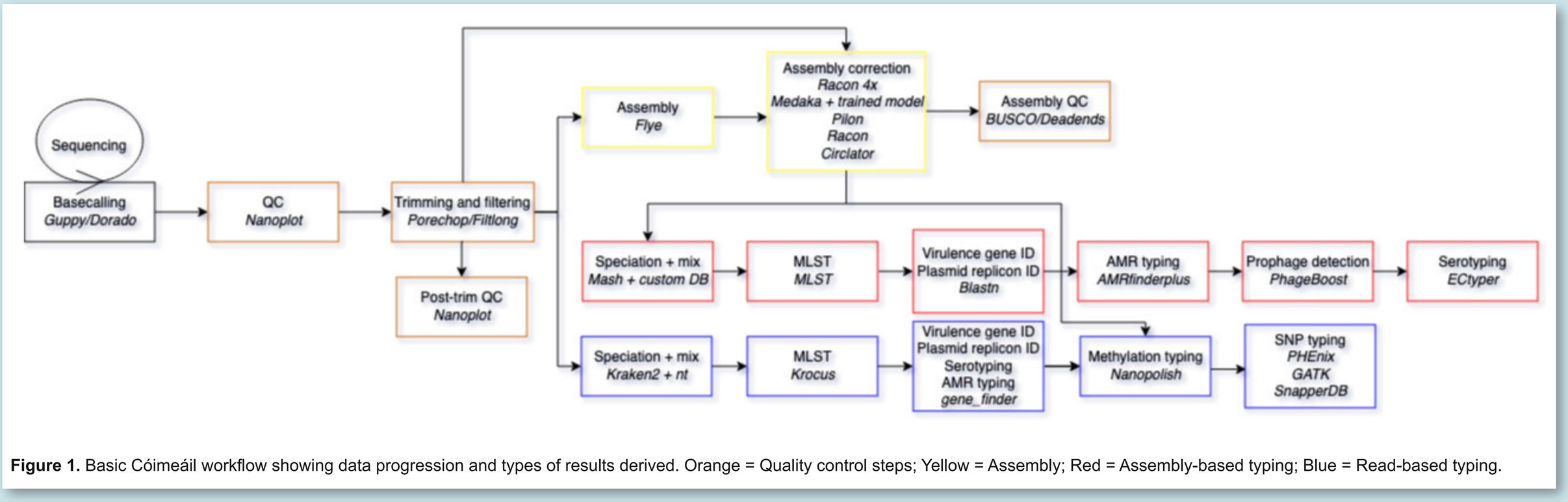


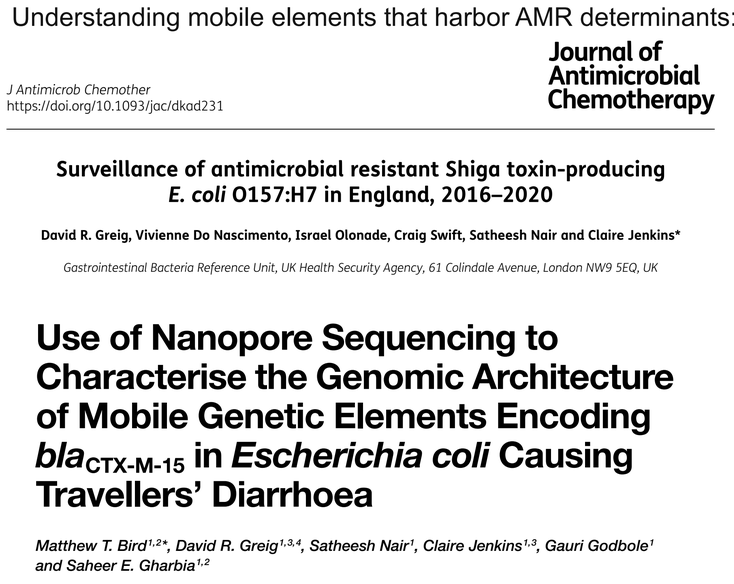
Figure 1. Basic Cóimeáil workflow showing data progression and types of results derived. Orange = Quality control steps; Yellow = Assembly; Red = Assembly-based typing; Blue = Read-based typing.

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].

WHERE HAS IT BEEN USED

- Understanding mobile elements that harbor AMR determinants:



J Antimicrob Chemother
https://doi.org/10.1093/jac/dkad231

Journal of Antimicrobial Chemotherapy

Surveillance of antimicrobial resistant Shiga toxin-producing *E. coli* O157:H7 in England, 2016–2020

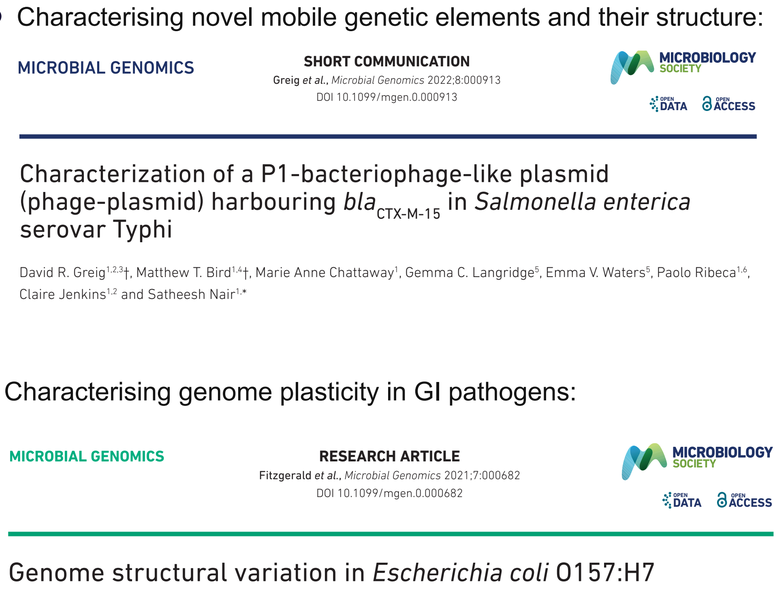
David R. Greig, Vivienne Do Nascimento, Israel Olonade, Craig Swift, Satheesh Nair and Claire Jenkins*

Gastrointestinal Bacteria Reference Unit, UK Health Security Agency, 61 Colindale Avenue, London NW9 5EQ, UK

Use of Nanopore Sequencing to Characterise the Genomic Architecture of Mobile Genetic Elements Encoding *bla*$_{CTX-M-15}$ in *Escherichia coli* Causing Travellers' Diarrhoea

Matthew T. Bird[1,2]*, David R. Greig[1,3,4], Satheesh Nair[1], Claire Jenkins[1,3], Gauri Godbole[1] and Saheer E. Gharbia[1,2]

- Characterising novel mobile genetic elements and their structure:



MICROBIAL GENOMICS

SHORT COMMUNICATION
Greig *et al.*, *Microbial Genomics* 2022;8:000913
DOI 10.1099/mgen.0.000913

MICROBIOLOGY SOCIETY
OPEN DATA OPEN ACCESS

Characterization of a P1-bacteriophage-like plasmid (phage-plasmid) harbouring *bla*$_{CTX-M-15}$ in *Salmonella enterica* serovar Typhi

David R. Greig[1,2,3†], Matthew T. Bird[1,4†], Marie Anne Chattaway[1], Gemma C. Langridge[5], Emma V. Waters[5], Paolo Ribeca[1,6], Claire Jenkins[1,2] and Satheesh Nair[1,*]

- Characterising genome plasticity in GI pathogens:

MICROBIAL GENOMICS

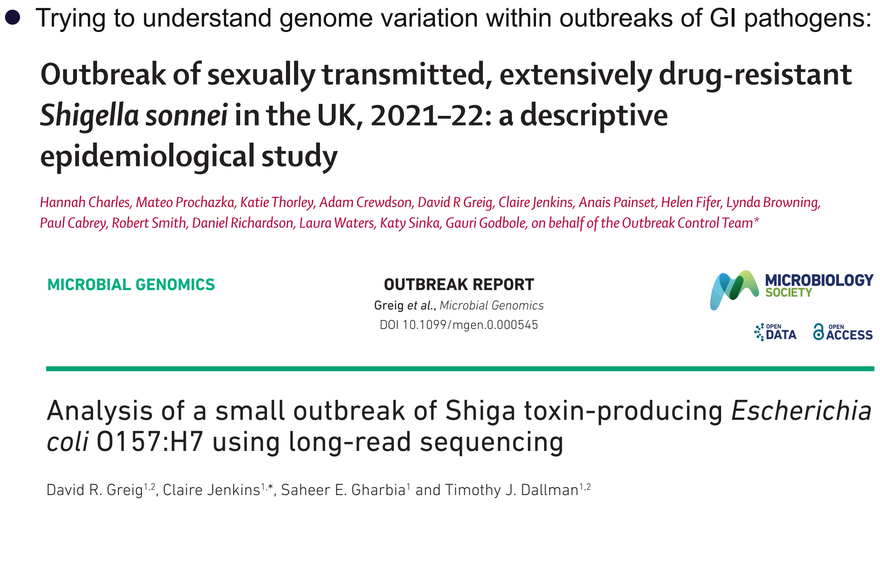
RESEARCH ARTICLE
Fitzgerald *et al.*, *Microbial Genomics* 2021;7:000682
DOI 10.1099/mgen.0.000682

MICROBIOLOGY SOCIETY
OPEN DATA OPEN ACCESS

Genome structural variation in *Escherichia coli* O157:H7

Stephen F. Fitzgerald[1,*], Nadejda Lupolova[1], Sharif Shaaban[1], Timothy J. Dallman[2], David Greig[2], Lesley Allison[3], Sue C. Tongue[4], Judith Evans[4], Madeleine K. Henry[4], Tom N. McNeilly[5], James L. Bono[6] and David L. Gally[1,*]

- Trying to understand genome variation within outbreaks of GI pathogens:



Outbreak of sexually transmitted, extensively drug-resistant *Shigella sonnei* in the UK, 2021–22: a descriptive epidemiological study

*Hannah Charles, Mateo Prochazka, Katie Thorley, Adam Crewdson, David R Greig, Claire Jenkins, Anais Painset, Helen Fifer, Lynda Browning, Paul Cabrey, Robert Smith, Daniel Richardson, Laura Waters, Katy Sinka, Gauri Godbole, on behalf of the Outbreak Control Team**

MICROBIAL GENOMICS

OUTBREAK REPORT
Greig *et al.*, *Microbial Genomics*
DOI 10.1099/mgen.0.000545

MICROBIOLOGY SOCIETY
OPEN DATA OPEN ACCESS

Analysis of a small outbreak of Shiga toxin-producing *Escherichia coli* O157:H7 using long-read sequencing

David R. Greig[1,2], Claire Jenkins[1,*], Saheer E. Gharbia[1] and Timothy J. Dallman[1,2]

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.
- Due to the modular nature of Cóimeáil, development of new components or an entire workflow for another gastrointestinal pathogen is possible with the long-term goal to be able to characterise all gastrointestinal pathogens that are processed by GBRU, UKHSA.
- Finally, work is underway to reformat the software under GBRU's services to try to fulfil the requirements for UKAS accreditation.

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.