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

@HannahTriv
hlhtrive@liverpool.ac.uk

Metagenomics For Pathogen Diagnostics: Problems Solved By Long Read Data

Hannah Trivett[1,2], Edward Cunningham-Oakes[1,2],Alistair C. Darby[1,2]

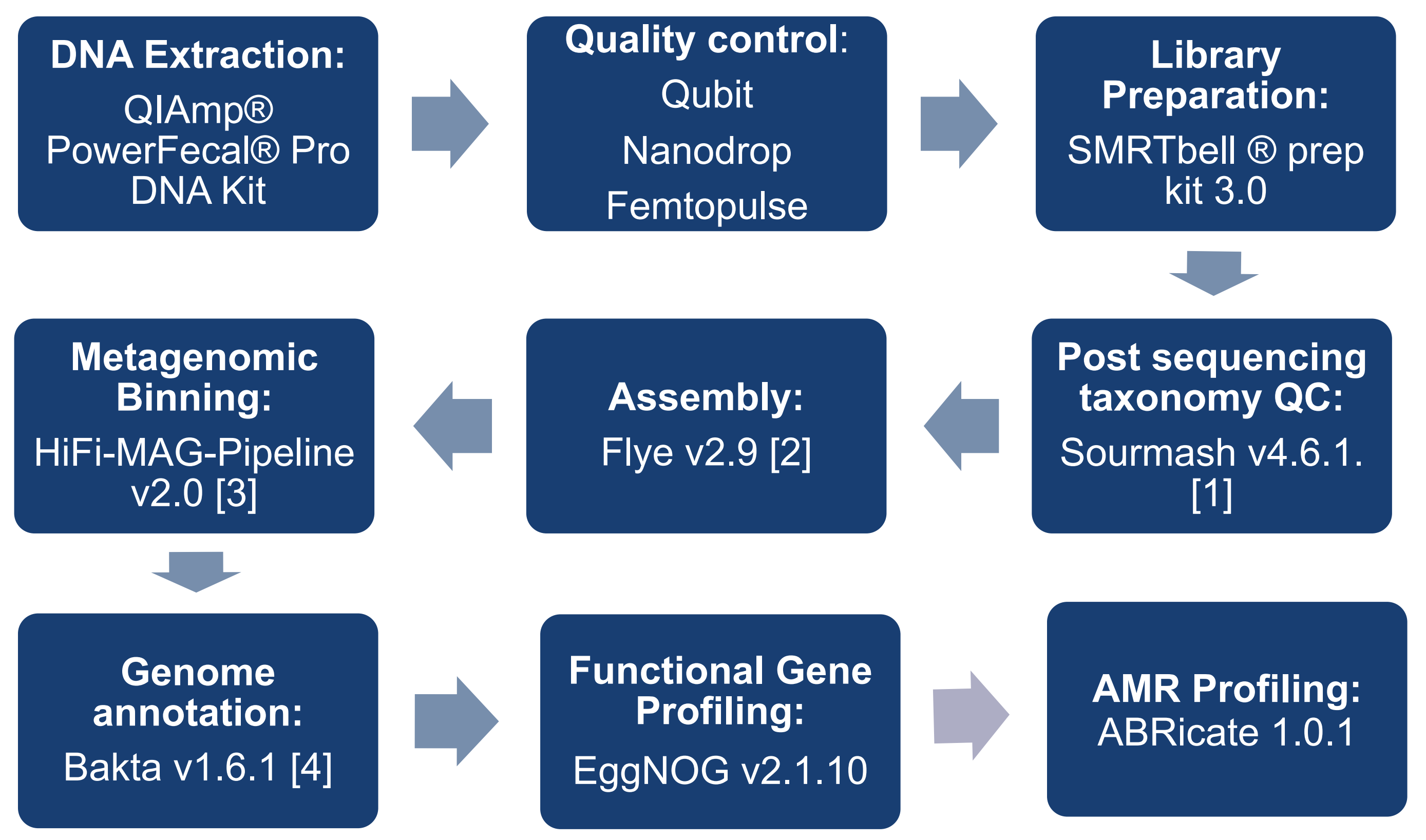
[1]Department of Infection Biology and Microbiomes, Institute of Infection, Veterinary and Ecological Sciences, University of Liverpool, Liverpool, UK.
[2]NIHR Health Protection Research Unit in Gastrointestinal Infections, Liverpool, UK.

Introduction:

The current gold-standard for profiling gut microbiomes is 16S ribosomal RNA-gene sequencing, or shotgun metagenomics using short-read sequencing data. However, developments in long-read sequencing accuracy have the potential to leverage new approaches for microbiome sequencing, providing a new standard.

Here we evaluate the effectiveness of PacBio HiFi long-reads to provide high quality metagenomic data, and crucial information, such as antimicrobial resistance profiles, from a single sequencing run. Our results demonstrate that PacBio HiFi long-read metagenomic sequencing shows promise for clinical applications as a culture-independent approach for rapid and accurate pathogen detection.

Methods:



DNA Extraction: QIAmp® PowerFecal® Pro DNA Kit

Quality control: Qubit, Nanodrop, Femtopulse

Library Preparation: SMRTbell ® prep kit 3.0

Post sequencing taxonomy QC: Sourmash v4.6.1. [1]

Assembly: Flye v2.9 [2]

Metagenomic Binning: HiFi-MAG-Pipeline v2.0 [3]

Genome annotation: Bakta v1.6.1 [4]

Functional Gene Profiling: EggNOG v2.1.10

AMR Profiling: ABRicate 1.0.1

Results

20 human stool samples were extracted using the Qiagen Power Fecal Pro HMW extraction kit. Of these samples, 13 possessed a gastrointestinal pathogen, as characterised using traditional and molecular diagnostics. Our aim was to understand and refine the use of read, contigs and MAG data to understand the diversity of taxa and genes present in the samples (Figure 1). We found that, across these 13 stool samples, the previously characterised pathogens were detectible in 13/13 HiFi reads, 11/13 contig assemblies and 5/13 of the high-quality metagenome-assembled genomes (MAGs) (Figure 1 and Table 1). MAGS corresponding to characterised pathogens were assembled as single contigs, highlighting the accuracy of PacBio HiFi sequencing technology. The focus on these high-quality data as MAGs potentially misses data for other taxa observed in the reads and contigs.

Sample ID	Pathogen ID	Number of Reads	Number of contigs	Number of MAGS >93% completeness	Number of taxa Reads	Number of taxa Contigs	Number of taxa MAGS	Pathogen present Reads	Pathogen present Contigs	Pathogen present MAGS
Sample 1	E.coli	10919743	3118	6	488	293	6	Yes	Yes	Yes
Sample 2	E.coli	9281284	8399	18	1237	578	18	Yes	Yes	No
Sample 3	E.coli	8076219	6760	8	1197	569	8	Yes	No	No
Sample 4	E.coli	12785965	6329	11	1055	546	11	Yes	Yes	No
Sample 5	E.coli	15130364	7554	19	1331	662	19	Yes	Yes	Yes
Sample 6	E.coli	13138948	1816	11	351	227	11	Yes	Yes	No
Sample 7	E.coli	14357474	6749	11	1135	545	11	Yes	Yes	No
Sample 8	E.coli	14973870	5196	14	922	451	14	Yes	Yes	No
Sample 9	E.coli	11096534	2799	11	487	261	11	Yes	Yes	Yes
Sample 10	Salmonella	2705815	341	5	102	36	5	Yes	Yes	Yes
Sample 11	Salmonella	27609987	432	6	191	88	6	Yes	Yes	Yes
Sample 12	Salmonella	44172522	8354	37	1582	946	37	Yes	No	No
Sample 13	N/A	4576822	5428	6	836	353	6	N/A	N/A	N/A
Sample 14	N/A	8283118	6950	8	902	448	8	N/A	N/A	N/A
Sample 15	N/A	7662029	10191	4	966	507	4	N/A	N/A	N/A
Sample 16	N/A	11042006	9119	11	1024	587	11	N/A	N/A	N/A
Sample 17	N/A	18589594	5251	16	1086	590	16	N/A	N/A	N/A
Sample 18	E.coli	10361223	6575	26	1302	712	26	Yes	Yes	No
Sample 19	N/A	12339003	4849	15	965	614	15	N/A	N/A	N/A
Sample 20	N/A	8935642	6319	17	1300	553	17	N/A	N/A	N/A

Table1: Summary statistics for 20 human stool samples, sequenced from the University of Liverpool BioBank and UKHSA. 13 samples tested positive for *Salmonella* or *E.coli* using the current clinical diagnostic frameworks. The number of taxa was calculated using Sourmash [4].

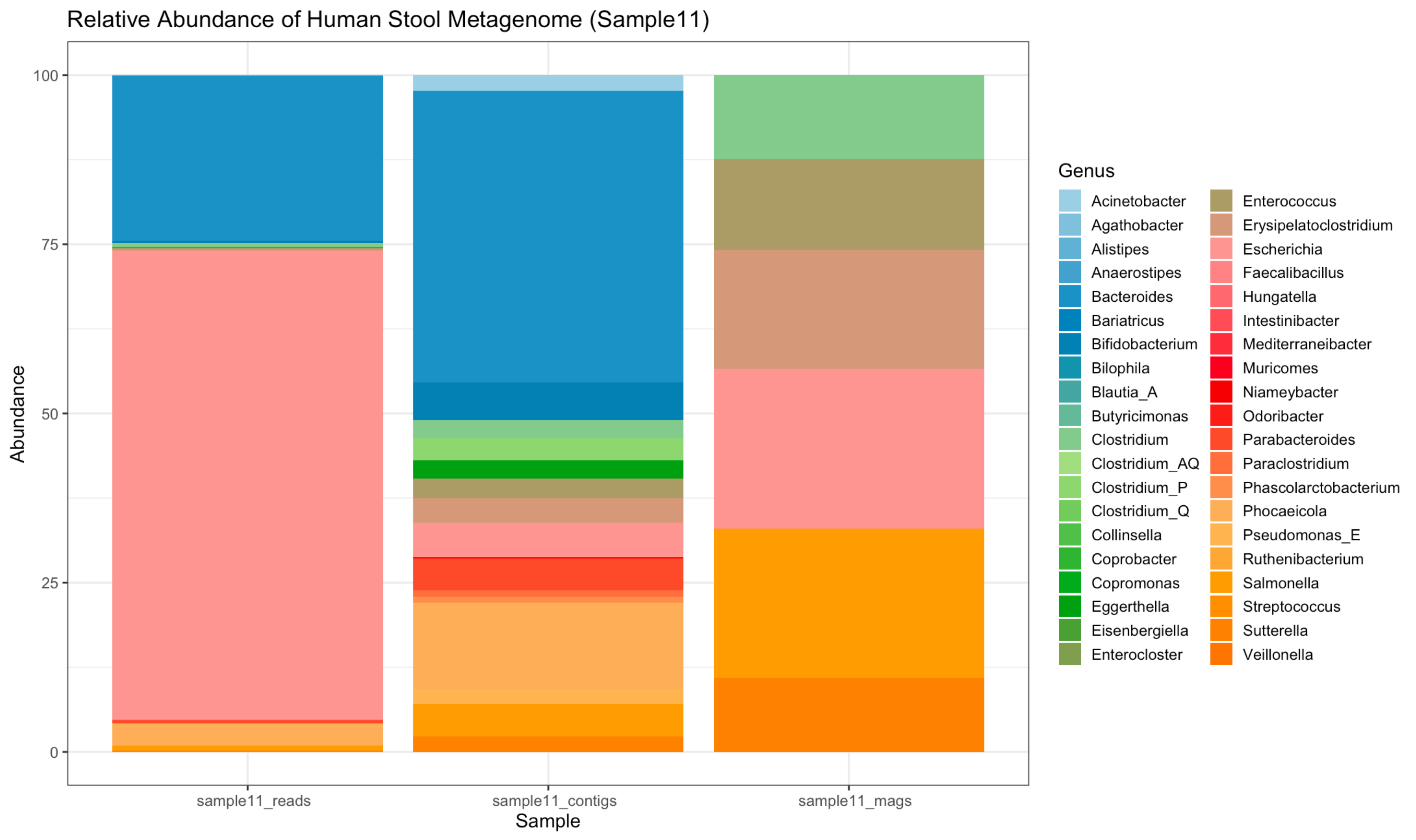


Figure 1: Comparison of the taxonomic composition relative abundance between reads, contigs and MAGs for sample 11. Taxonomic classification was completed using Sourmash, using GTDB-R207 as a reference database.

PacBio HiFi metagenomic data was used to generate a *Salmonella* MAG, which was scored as 100% complete by CheckM2. This corroborated with previous traditional and molecular results. The isolate and MAG genomes are comparable, with sequence lengths of 4679226 and 4679254, respectively (Figure 2). In addition, both genomes have 4334 protein-coding genes and a GC content of 52.2%, highlighting the accuracy of long-read metagenomic sequencing to produce isolate-quality MAG sequences to characterise genomes of interest.

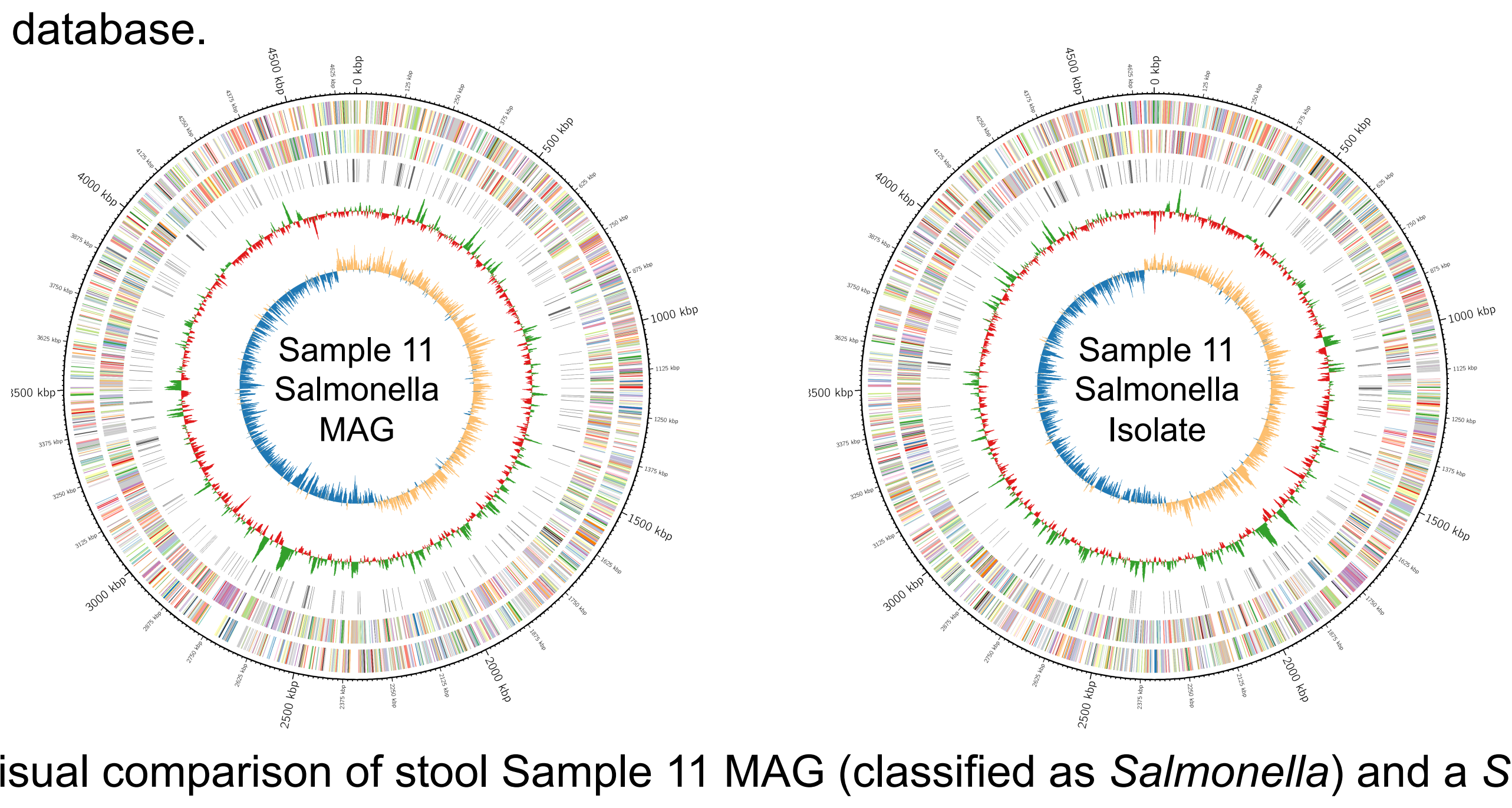


Figure 2: Visual comparison of stool Sample 11 MAG (classified as *Salmonella*) and a *Salmonella* genome, from a cultured isolate, derived from the same stool. Circos plots were produced by Bakta.

Conclusions and future directions:

PacBio HiFi long-read sequencing was able to provide a clinically-relevant characterisation of the human gut microbiome, with pathogen genome identification possible at the species level. This was comparable to the resolution provided by bacterial isolate sequencing.
Future work will compare PacBio HiFi long-read metagenomic data with Illumina paired-end short-read data. We will also compare genomic variation between long and short-read isolate contigs derived from the same stool sample.

References:

[1] https://github.com/PacificBiosciences/pb-metagenomics-tools/blob/master/docs/Tutorial-HiFi-MAG-Pipeline.md.
[2] Kolmogorov, *et al.* (2020). "metaFlye: scalable long-read metagenome assembly using repeat graphs", *Nature Methods*, 17(1). Available at: https://doi.org/10.1038/s41592-020-00971-x.
[3] Titus Brown, C. and Irber, L. (2016). "Sourmash: A library for Minhash sketching of DNA" *The Journal of Open Source Software*, 1(5). Available at: https://doi.org/10.21105/joss.00027.
[4] Schwengers O *et al.* (2021). "Bakta: rapid and standardized annotation of bacterial genomes via alignment-free sequence identification" *Microbial Genomics*, 7(11). Available at: https://doi.org/10.1099/mgen.0.000685.

UNIVERSITY OF LIVERPOOL

WARWICK
THE UNIVERSITY OF WARWICK

UK Health Security Agency

FUNDED BY
NIHR | National Institute for Health and Care Research

The research was funded by the National Institute for Health and Care Research Health Protection Research Unit (NIHR HPRU) in Gastrointestinal Infections at University of Liverpool in partnership with the UK Health Security Agency (UKHSA), 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 the UK Health Security Agency.