**A pangenome analysis comparing and contrasting *Salmonella enterica* subspecies *enterica* and *Salmonella enterica* subspecies *salamae* isolated from households in Malawi.**

Catherine N. Wilson\*, Mathew A. Beale, Patrick Musicha, Yohane Diness, Oscar Kanjerwa, Nicholas Feasey, Eric Fèvre, Gina Pinchbeck, Nicholas Thomson

**Methods: Genomic processing**

***Quality control***

FastQC (version 0.11.9) and multiQC (version 0.11.8) were used to assess per base sequence quality, quality scores per sequence, per base sequence content, per base GC content, per sequence GC content, per base N content, contig length distribution and sequence duplication levels (1).

Read quality control was undertaken using Kraken (version 1.1.1) to assess the relative abundance of different bacteria within the reads by assigning taxonomic levels to short DNA sequences (2). CheckM (version 1.1.2) was run to assess contamination, strain heterogeneity and completeness of the genomes (3). Assembly Statistics (version 1.0.1) was run to analyse the total genome length and number of contigs (4). The Quality assessment tool for genome assemblies (QUAST) (version 5.0.2) was used to assess the number of contigs, N50 and total length of the genome (5). Following the completion of quality control procedures, genomes were submitted to Pathogenwatch, which uses SISTR to assess the species, serovar and sequence type of the bacteria present (6,7). SPADES (version 3.14) was used to assemble the genomes and Prokka (version 1.14.5) was used for genome annotation (8,9).

***Core-genome phylogeny and SNP analysis***

A core and pangenome analysis was performed using Roary (version 3.11.2)(10). A gene was considered core if it was present in 100% of the genomes at a match identity threshold of 95%. A core genome sequence alignment was generated using Roary by concatenating the alignments of the core genes. Single nucleotide polymorphic (SNP) site alignment was generated from the core genome alignment using SNP-sites (version 2.5.1) (11). RAxML (Randomised Accelerated Maximum Likelihood)(version 8.2.8) was run on the resulting core SNP-alignment to construct a maximum likelihood tree using the core gene SNP alignment of all 227 isolates (12). Reliability of inferred branch partitions was assessed with 100 bootstrap replicates. The tree was visualised using ITOL (version 5) and ggtree (version 3.2) (13,14). Panstripe was used to examine the pangenome of the 227 isolates (15).

#### 

#### **Identification of AMR determinants, virulence factors and plasmid typing**

AMRFinderPlus (version 3.10) was used to detect chromosomal mutations encoding for AMR, acquired AMR genes (ARGs) and heavy metal resistance genes (16). This programme uses a BLAST approach to identify AMR genes. Those ARGs with an identity of 95% and a coverage of 95% were taken forward for further analysis. ABRIcate was also used to search for virulence genes, using the VFDB (Virulence Factor Database)(17). Those virulence factors with greater than or equal to 95% coverage and 95% identity were taken forward for further analysis. Plasmid analysis of the 227 study assemblies was carried out using MOB\_recon, a software tool of MOB\_suite (18). ABRIcate was used to search the plasmid and chromosomal contigs for AMR determinants and virulence genes using the VFDB database(17,19).

**References**

1. Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data [Internet]. [cited 2021 Nov 17]. Available from: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

2. Kraken: ultrafast metagenomic sequence classification using exact alignments | Genome Biology | Full Text [Internet]. [cited 2021 Nov 17]. Available from: https://genomebiology.biomedcentral.com/articles/10.1186/gb-2014-15-3-r46

3. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. 2015 Jul;25(7):1043–55.

4. assembly-stats [Internet]. Pathogen Informatics, Wellcome Sanger Institute; 2022 [cited 2022 Jul 16]. Available from: https://github.com/sanger-pathogens/assembly-stats

5. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinforma Oxf Engl. 2013 Apr 15;29(8):1072–5.

6. Pathogenwatch | A Global Platform for Genomic Surveillance [Internet]. [cited 2021 Nov 17]. Available from: https://pathogen.watch/

7. sistr\_cmd: Salmonella In Silico Typing Resource (SISTR) commandline tool [Internet]. National Microbiology Laboratory; 2022 [cited 2022 Aug 5]. Available from: https://github.com/phac-nml/sistr\_cmd

8. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. J Comput Biol. 2012 May;19(5):455–77.

9. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinforma Oxf Engl. 2014 Jul 15;30(14):2068–9.

10. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MTG, et al. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics. 2015 Nov 15;31(22):3691–3.

11. Page AJ, Taylor B, Delaney AJ, Soares J, Seemann T, Keane JA, et al. SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. Microb Genomics. 2(4):e000056.

12. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014 May 1;30(9):1312–3.

13. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. Nucleic Acids Res. 2021 Jul 2;49(W1):W293–6.

14. Yu G. Using ggtree to Visualize Data on Tree-Like Structures. Curr Protoc Bioinforma. 2020;69(1):e96.

15. Tonkin-Hill G, Gladstone RA, Pöntinen AK, Arredondo-Alonso S, Bentley SD, Corander J. Robust analysis of prokaryotic pangenome gene gain and loss rates with Panstripe [Internet]. bioRxiv; 2022 [cited 2022 Jul 12]. p. 2022.04.23.489244. Available from: https://www.biorxiv.org/content/10.1101/2022.04.23.489244v1

16. AMRFinderPlus - Pathogen Detection - NCBI [Internet]. [cited 2021 Nov 17]. Available from: https://www.ncbi.nlm.nih.gov/pathogens/antimicrobial-resistance/AMRFinder/

17. Seemann T. ABRicate [Internet]. 2021 [cited 2021 Nov 17]. Available from: https://github.com/tseemann/abricate

18. Robertson J, Nash JHE. MOB-suite: software tools for clustering, reconstruction and typing of plasmids from draft assemblies. Microb Genomics. 2018 Aug;4(8).

19. Chen L, Yang J, Yu J, Yao Z, Sun L, Shen Y, et al. VFDB: a reference database for bacterial virulence factors. Nucleic Acids Res. 2005 Jan 1;33(Database Issue):D325–8.