Improving pathogen analysis by pushing the boundaries of long read sequencing

Steven John Rudder
0418188
Quadram Institute
University of East Anglia

This thesis is submitted for the degree of Doctor of Philosophy
August 2025

Word count: 66,012

The author, S. J. Rudder, drafted all original text of this thesis. ChatGPT-4 was employed in a limited capacity to enhance grammar and readability.

Abstract

Salmonella and *Campylobacter* are leading foodborne pathogens responsible for gastroenteritis globally, yet their detection and characterisation remain limited by culturing challenges, DNA extraction constraints, and preservation-related biases. Advances in long read sequencing platforms and metagenomic approaches offer exciting opportunities to overcome these barriers by enabling culture-free recovery of complete genomes directly from stool.

This work combined laboratory automation for high molecular weight (HMW) DNA extraction with short- and long- read sequencing to address four key challenges: (i) development of semi-automated Fire Monkey protocols for HMW DNA extraction on a Tecan A200 robotic platform for clinical bacterial isolates and stool, (ii) investigation of within-host diversity of *Salmonella enterica* from gastroenteritis patients, (iii) evaluation of stool preservation conditions for metagenomic recovery of *Campylobacter* genomes, and (iv) implementation of HMW-DNA extraction and long-read sequencing from stool for metagenomic recovery of *Campylobacter* genomes.

Developed Fire Monkey protocols produced DNA of sufficient length and purity for long-read sequencing and hybrid assembly. This enabled single contig bacterial genome to be assembled with DNA extracted from both isolates and stool. Sequencing of up to 20 *Salmonella* colonies per patient revealed within-host diversity was limited to single nucleotide polymorphisms (SNPs) and antimicrobial resistance gene profiles. In the *Campylobacter* storage study, stool frozen untreated and stool frozen with glycerol outperformed Zymo DNA/RNA Shield for preserving genome coverage and typing accuracy over nine months at –80 °C. The Fire Monkey stool HMW DNA protocol developed as part of the project enabled recovery of a single-contig *Campylobacter* genome, which facilitated typing at SNP resolution. Comparative evaluation of Fire Monkey and Maxwell extractions further demonstrated that DNA quality strongly influenced the completeness and reliability of metagenome-derived genomes.

Together, these findings help to inform best practices for public health surveillance, outbreak investigations, and the future integration of metagenomics.

Acknowledgements

I lovingly dedicate this thesis to the memory of my father, Graham Rudder. Thank you for showing me the beauty of the natural world. As a child, walking among giant sequoias, exploring mountains, deserts, and canyons instilled in me a deep sense of wonder and a lifelong desire to explore the natural world.

A huge thank you to my primary supervisor Dr. Gemma Langridge for advice, reviewing all my work, and keeping me on track. Thank you to Dr. Nicol Janecko for supporting and reviewing stool-based studies. I'd also like to thank advisory panel members Prof. Mark Webber and Prof. John Wain for taking part in discussion throughout the project.

Some shout outs to key members and influences throughout the journey. Dr. Emma Waters for offering the time and energy to support in all areas of the project from lab work, through discussion, and proof-reading. Dr. Bilal Djeghout generously shared his expertise in stool DNA extraction and metagenomics, providing me with a solid foundation on which to build. Alice Nisbet, it's been an honour to sit next to you for three years, thanks for putting up with all my questions and helping we command line prompts. Dr. Heather Felgate for supplying coffee beans and all sort of wisdom scientific and beyond. Finally, Dave Baker, before I started my PhD, you turned what could have been a monotonous role in the sequencing department into an exciting place to work through your passion for innovation and commitment to pushing the technology forward.

Table of Contents

1 Introduction 19

1.1 The Role of Genomics in Outbreaks and Antimicrobial Resistance 19
1.1.1 Outbreak Investigation and Surveillance 19
1.1.2 Antimicrobial Resistance Monitoring 20
1.1.3 Overview of Foodborne Pathogens: *Salmonella* and *Campylobacter* 21
1.1.4 Brief History of Sequencing Technologies in Epidemiology 24

1.2 Short-Read Sequencing in Public Health Applications 25
1.2.1 Advantages of Short-Read Sequencing 25
1.2.2 Limitations of Short-Read Sequencing 26
1.2.3 Applications in Public Health Genomics 27

1.3 High-Throughput Sequencing and Metagenomic Approaches in Pathogen and Community Analysis 31
1.3.1 High-Throughput Sequencing Pipelines and Their Scalability 31
1.3.2 Metagenomics for Stool Samples and Microbial Communities 33
1.3.3 Applications in Transmission Dynamics and Outbreak Source Tracking 34

1.4 Advancing Pathogen Analysis with Long-Read Sequencing 35
1.4.1 Overcoming the Limitations of Short-Read Sequencing 36
1.4.2 Error Correction and Assembly Polishing 36
1.4.3 Applications in Long Read Sequencing Enteric Pathogen Surveillance 37
1.4.4 Future Outlook 39

1.5 DNA Extraction and Sequencing Workflow Optimisation for Pathogen Analysis 39
1.5.1 Challenges in DNA Extraction from Different Sample Types 39
1.5.2 High-Quality DNA Yield for Short-Read vs Long-Read Sequencing 42
1.5.3 Best Practices for Data Quality and Reproducibility 44
1.5.4 Speed and Cost Considerations in Workflow Optimisation 46

1.6 Aims, Objectives, & Research Questions 49

2 Automating High Molecular Weight DNA Extraction: Fire Monkey Protocols for bacterial isolates and stool on the Tecan A200 52

2.1 Introduction 52

2.2 Aims and objectives 57

2.3 Materials and methods 57
2.3.1 DNA quantification – Single tube assay 57
2.3.2 DNA quantification – Plate assay 57
2.3.3 DNA sizing 58
2.3.4 Nanodrop 58
2.3.5 Proteinase K 58
2.3.6 Host depletion reagents 58
2.3.7 Bead clean 58
2.3.8 Preparing Polyvinylpolypyrrolidone 59
2.3.9 Stool collection 59
2.3.10 Fire Monkey using the A200 60

2.4 Results 63
2.4.1 Testing the Fire Monkey Tecan A200 platform with *Escherichia coli* 63
2.4.2 Developing a Fire Monkey Tecan A200 protocol for *Salmonella* 66
2.4.3 Developing a Fire Monkey Tecan A200 protocol for stool 73

2.5 Discussion 99
2.5.1 HMW Stool DNA Extraction with Fire Monkey 99
2.5.2 HMW bacterial cell DNA Extraction with Fire Monkey 103
2.6 Conclusion 104
3 Genomic diversity of non-typhoidal *Salmonella* found within patients suffering from gastroenteritis in Norfolk, UK 105
3.1 Introduction 105
3.2 Aims and objectives 107
3.3 Materials and methods 107
3.3.1 Stool collection 107
3.3.2 Bacterial isolation 108
3.3.3 DNA extraction 108
3.3.4 DNA quantification 109
3.3.5 DNA cleaning & concentrating 111
3.3.6 Bacterial isolate short-read sequencing 111
3.3.7 Bacterial isolate long-read sequencing 112
3.3.8 Genome assembly 114
3.3.9 Genome assembly quality control 117
3.3.10 Genome annotation 117
3.3.11 *In silico* typing and AMR predictions 117
3.3.12 Genome structural analysis 117
3.3.13 Single nucleotide polymorphism analysis 118
3.3.14 Sequence alignment, read mapping and visualisation 118
3.3.15 Hierarchical clustering 118
3.4 Results 118
3.4.1 Stool specimen and linked metadata 118
3.4.2 Bacterial isolation 119
3.4.3 DNA sequencing 123
3.4.4 Issues with Hybrid Assemblies for SNP Analysis 123
3.4.5 Genome Level Diversity – Serovar and Sequence Type 138
3.4.6 Genome Level Diversity – Antimicrobial Resistance Determinants 138
3.4.7 Genome Level Diversity – Genome Structure Analysis 139
3.4.8 Genome Level Diversity – SNP Analysis 139
3.4.9 Hierarchical clustering of *Salmonella* Java 145
3.5 Discussion 149
3.5.1 Hybrid assembly 150
3.5.2 Genome Level Diversity 151
3.6 Conclusions 158
4 Impact of Preservation Conditions on the Recovery of Metagenome Derived *Campylobacter* genomes from Stool Samples 160
4.1 Introduction 160
4.2 Aims and objectives 163
4.3 Methods 164
4.3.1 Experimental design overview 164
4.3.2 Sample collection 164
4.3.3 Bacterial isolation 167
4.3.4 Stool sample preservation conditions and storage 167
4.3.5 DNA extraction from stool 167
4.3.6 DNA extraction from isolates 169

4.3.7 DNA quantification 169
4.3.8 Quantitative qPCR 170
4.3.9 DNA sequencing library preparation 170
4.3.10 DNA sequencing 171
4.3.11 *in-silico* human read removal 172
4.3.12 Bacterial isolate assembly 172
4.3.13 *Campylobacter* read recovery from metagenome sequencing 172
4.3.14 Recovered read assembly 172
4.3.15 Classification 172
4.3.16 Multi-Locus Sequencing Typing 173
4.3.17 Antimicrobial resistance genotyping 173
4.3.18 Read mapping 173
4.3.19 Genome assembly quality assessment by QUAST 174
4.3.20 CheckM analysis of metagenome derived genome (MDG) completeness 174
4.3.21 Statistics 174

4.4 Results 175
4.4.1 Overview of stool samples 175
4.4.2 *Campylobacter* isolation 177
4.4.3 Sequence typing for metagenome derived genomes 179
4.4.4 Classification 180
4.4.5 Antimicrobial resistance genotypes in metagenome-derived *Campylobacter* genomes 182
4.4.6 Statistical analysis of storage conditions using coverage scores 183
4.4.7 qPCR 199
4.4.8 N50 202
4.4.9 Human Host DNA contamination 203
4.4.10 CheckM completeness 204

4.5 Discussion 210
4.5.1 Effect of three preservation conditions and -80°C storage for up to 9-months on *Campylobacter* detection and typing 210
4.5.2 Implications for Diagnostic Laboratories and Resource-Limited Settings 215

4.6 Conclusion 218

5 Detection of *Campylobacter* with long-read sequencing of DNA from human stool 219

5.1 Introduction 219

5.2 Aims and objectives 221

5.3 Methods 221
5.3.1 Sample collection 221
5.3.2 DNA extraction from stool 221
5.3.3 DNA extraction from isolates 222
5.3.4 DNA sequencing – Metagenome 222
5.3.5 DNA sequencing – Isolates 224
5.3.6 Long-read metagenome assembly pipeline 224
5.3.7 Short-read metagenome assembly pipeline 224
5.3.8 *Campylobacter* bin identification 224
5.3.9 Kraken read recovery MAGs 225
5.3.10 Isolate assembly 225
5.3.11 Typing and Antimicrobial resistance determinant identification 225
5.3.12 Single Nucleotide Polymorphism analysis 225
5.3.13 Relative abundance 226
5.3.14 Read mapping 226

5.4 Results 226

5.4.1 Sample information 226
5.4.2 Stool DNA MinION runs 226
5.4.3 Long-read size and quality filtering 230
5.4.4 Basic fasta statistical comparison of long- and short-read *Campylobacter* MAGs 232
5.4.5 MAG completeness 236
5.4.6 *Campylobacter* isolates 238
5.4.7 Mapping metagenomic reads to isolates 238
5.4.8 MAGs versus isolate typing 239
5.4.9 MAGs versus isolate antimicrobial determinants 242
5.4.10 Tetracycline resistance determinants 244
5.4.11 MAGs versus isolate single nucleotide polymorphisms 245
5.4.12 Community composition 248
5.4.13 qPCR 253

5.5 Discussion 254

5.6 Conclusion 259

6 General Discussion 260

6.1 Methodological advancements 260

6.2 Within-host genomic diversity of *Salmonella* 262

6.3 Impact of stool preservation on *Campylobacter* DNA 263

6.4 Direct recovery of *Campylobacter* genomes from stool with long reads 263

6.5 Challenges facing stool metagenomics as a pathogen detection tool . 264

6.6 Public health implications and future outlook 264

6.7 Future directions 265
6.7.1 General 266
6.7.2 Focused 266

6.8 Final remarks 267

7 References 269

Appendices

Appendix 1 - Key Developmental Protocol Variants **298**

Appendix 2 - Sequencing stats for metagenome samples used in Chapter 4. Stats include raw data yield, reads_in represents total read yield and reads_out represents the number of reads after human read removal. **301**

Appendix 3 - AMR in-silico predictions for isolates and MD-*Campylobacter* genomes used in Chapter 4 ..... **305**

Appendix 4 - Raw and normalised data input for statistical tests, MD-*Campylobacter* genomes used in Chapter 4 ..... **318**

Appendix 5 - Full Shapiro-Wilk test for coverage metrics **324**

Appendix 6 - MD-*Campylobacter* genomes GTDB-Tk classification, and mean qPCR results used in Chapter 4 ..... **326**

Appendix 7 - Isolate and MD-*Campylobacter* genomes full MLST scores used in Chapter 4 ..... **331**

Appendix 8 - CheckM results for MD-*Campylobacter* genomes, these values are not standardised to reads in and represent the full sequencing yield of each sample ... **342**

List of Tables

Table 2.1: Approximations of DNA Yield and Fragment Size of Commercial HMW DNA Extraction Kits Targeting Bacterial Cells 55
Table 2.2: Approximations of DNA Yield and Fragment Size of Commercial HMW DNA Extraction Kits Targeting Human Stool 56
Table 2.3: Template Protocol for Fire Monkey HMW DNA Extraction Kit Washes and Elution 60
Table 2.4: Tecan A200 operations for *E. coli* SR version 2 protocol 64
Table 2.5: Tecan A200 Operations for *Salmonella* Protocol 68
Table 2.6: CTAB Extraction Experiment Protocol Variant with Stool Input Weight and Resulting DNA Yield 76
Table 2.7: Enzymatic Digestion Experiment Protocol Variant with Stool Input Weight and Resulting DNA Yield 77
Table: 2.8 Addition of Stool Washing Experiment Protocol Variant with Stool Input Weight and Resulting DNA Yield Plus DNA Integrity Number (DIN) 81
Table 2.9: Addition of Alcohol Washing Steps Experiment Protocol Variant with Stool Input Weight and Resulting DNA Yield Plus DNA Integrity Number (DIN) 83
Table 2.10: Addition of Neutrase and Increasing Reagent Volume Experiment Protocol Variant with Stool Input Weight and Resulting DNA Yield Plus DNA Integrity Number (DIN) 87
Table 2.11: Max-RSC Versus FM-W-3x Experiment Protocol Variant with Stool Input Weight and Resulting DNA Yield 90
Table 2.12: Nanodrop Values for Stool DNA Extracts Pre- and Post-SPRI Bead Clean 91
Table 2.13: Nanodrop Values for Stool DNA Extracts From Final Cleaning Testing 93
Table 2.14: Nanodrop Values for SPRI Bead Cleaned DNA From Final Cleaning Testing 94
Table 2.15: Nanodrop Values Pre- and Post-SPRI Bead Clean for Final Adjustment Testing 95
Table 2.16: Tecan A200 Stool Protocol 96
Table 3.1: Time Intervals and Bacterial Isolation Data for *Salmonella enterica* Subspecies from Patients 120
Table 3.2: Sample Origin and Patient Metadata 121
Table 3.3: Isolation of Salmonella from Stool Samples by Media Type Used for Colony Selection 122
Table 3.4: Number of Isolates with Sequencing Files Cleared for Assembly 123
Table 3.5: Overview of SNP Calls from 22EPA051NSA LR-Pilon Hybrid Assemblies 129
Table 3.6: Estimated Read Coverage of 20EPA002SNSA Isolate Genomes and SNP Calls for Different Assembly Pipelines 130
Table 3.7: Estimated Read Coverage of 20EPA011SNSA Isolate Genomes and SNP Calls for Different Assembly Pipelines 131
Table 3.8: Estimated Read Coverage of 20EPA012SNSA Isolate Genomes and SNP Calls for Different Assembly Pipelines 132
Table 3.9: Estimated Read Coverage of 22EPA044SNSA Isolate Genomes and SNP Calls for Different Assembly Pipelines 133
Table 3.10: Estimated Read Coverage of 22EPA051SNSA Isolate Genomes and SNP Calls for Different Assembly Pipelines 134

Table 3.11: Estimated Read Coverage of 22EPA053SNSA Isolate Genomes and SNP Calls for Different Assembly Pipelines 135
Table 3.12: Estimated Read Coverage of 22EPA055SNSA Isolate Genomes and SNP Calls for Different Assembly Pipelines 136
Table 3.13: Estimated Read Coverage of 22EPA058SNSA Isolate Genomes and SNP Calls for Different Assembly Pipelines 137
Table 3.14: *Salmonella* Classification for Each Stool Specimen 138
Table 3.15: Summary of Genetic Variants Identified in 20EPA002NSA Isolates... 140
Table 3.16: Summary of Genetic Variants Identified in 20EPA011NSA Isolates... 142
Table 3.17: Summary of Genetic Variants Identified in 20EPA012NSA Isolates... 143
Table 3.18: Summary of Genetic Variants Identified in 22EPA051NSA Isolates... 144
Table 3.19: Summary of Genetic Variants Identified in 22EPA055NSA Isolates... 144
Table 3.20: Summary of Genetic Variants Identified in 22EPA058NSA Isolates... 145
Table 4.1: Summary of Stool Samples in the Storage Conditions Experiment 176
Table 4.2: Classification, Sequence Type (ST), and AMR Determinants Identified in *Campylobacter* Isolates from Stool Samples 178
Table 4.3: MLST Allele Score for Metagenome Derived Genomes Versus Isolate Reference Genome, 7 Alleles Represent a Complete MLST Profile Resulting in a Sequence Type 180
Table 4.4: GTDB-Tk Classification of Metagenome Derived Genomes Compared to Isolate References 181
Table 4.5: Number of AMR Determinants Correctly Identified in Metagenome Derived *Campylobacter* Genomes Versus Isolate References for Each Stool and Storage Condition 182
Table 4.6: Overview of Shapiro-Wilk Tests for Normality of the Distribution of Genome Coverage Scores 184
Table 4.7: Wilcoxon Test Results for Breadth, Depth, and Genome Fraction of MD-*Campylobacter* Genomes from Stool Stored in Different Conditions from 0-9 months, Normalised to "Reads in" and Reported per 10 Million Reads 185
Table 4.8: Wilcoxon Test Results for Breadth, Depth, and Genome fraction, of MD-*Campylobacter* Genomes from Stool Stored in Different Conditions from 0-9 Months, with log10 Transformation Applied to Values 186
Table 4.9: Wilcoxon signed-rank tests timepoints 1 vs 9 198
Table 4.10: Wilcoxon signed-rank tests timepoints 1 vs 3 198
Table 4.11: Wilcoxon signed-rank tests timepoints 3 vs 9 199
Table 4.12: Comparison of multivariable logistic regression model with and without 130 and 135 replicates 202
Table 5.1: DNA Quantifications and Nanodrop Ratios Assessing Purity of Nucleic Acids 227
Table 5.2: Final Nanopore Library for Loading on the MinION 227
Table 5.3: Basic MinION Run Information 230
Table 5.4: Read Statistics for Stool 164 Fire Monkey DNA Prep 231
Table 5.5: Read Statistics for Stool 164 Maxwell DNA Prep 231
Table 5.6: Read Statistics for Stool 165 Fire Monkey DNA Prep 231
Table 5.7: Read Statistics for Stool 165 Maxwell DNA Prep 231
Table 5.8: Fasta Statistics on Stool 164 MDGs Assembled Using the Different Pipelines 234
Table 5.9: Fasta statistics on Stool 165 MDGs assembled using the different pipelines 235
Table 5.10: CheckM Results on MetaMBDG *Campylobacter* Bins Using Marker Lineage *Campylobacter* (UID3076) 237

Table 5.11: CheckM Results on MetaWrap2 *Campylobacter* Bins Using Marker Lineage *Campylobacter* (UID3076) 237
Table 5.12: CheckM Results on *Campylobacter* Reads Recovered Using Kraken2 238
Table 5.13: Mapping of Metagenome Sequencing Long Reads to *Campylobacter* Isolate 239
Table 5.14: Mapping of Metagenome Sequencing Short Reads to *Campylobacter* Isolate 239
Table 5.15: MLST Results for Stool 164 for Fire Monkey and Maxwell Preps Sequenced by MinION (Long) and Illumina Pair-end 150 bp (Short) with Assemblies Created by Binning and Read Classification Approaches 240
Table 5.16: MLST Results for Stool 165 for Fire Monkey and Maxwell Preps Sequenced by MinION (Long) and Illumina Pair-end 150 bp (Short) with Assemblies Created by Binning and Read Classification Approaches 241
Table 5.17: AMR Detection Results for Stool 164 for Fire Monkey and Maxwell Preps Sequenced by MinION (Long) and Illumina Pair-end 150 bp (Short) with Assemblies Created by Binning and Read Classification Approaches 242
Table 5.18: AMR Detection Results for Stool 165 for Fire Monkey and Maxwell Preps Sequenced by MinION (Long) and Illumina Pair-end 150 bp (Short) with Assemblies Created by Binning and Read Classification Approaches 243
Table 5.19: Manual and AbritAMR Identification of *tet*(*O*) in MDGs and Isolate Sequencing 245
Table 5.20: Single Nucleotide Polymorphism Analysis Results for Stool 164 for Fire Monkey and Maxwell Preps Sequenced by MinION (Long) and Illumina Pair-end 150 bp (Short) with Assemblies Created by Binning and Read Classification Approaches 246
Table 5.21: Single Nucleotide Polymorphism Analysis Results for Stool 165 for Fire Monkey and Maxwell Preps Sequenced by MinION (Long) and Illumina Pair-end 150 bp (Short) with Assemblies Created by Binning and Read Classification Approaches 247
Table 5.22: Relative Abundance of the Top 15 Genera for Stool 164 for Fire Monkey and Maxwell Preps Sequenced by MinION (Long) and Illumina Pair-end 150 bp (Short) 248
Table 5.23: Relative Abundance of the Top 15 Genera for Stool 165 for Fire Monkey and Maxwell Preps Sequenced by MinION (Long) and Illumina Pair-end 150 bp (Short) After Removal of *Homo* genus 253
Table 5.24: *POLR2A* qPCR Values for Stool DNA Preps and Human Control DNA 253
Table 5.25: *cadF* qPCR Values for Stool DNA Preps and *Campylobacter* Control DNA 254

Table of Figures

Figure 2.1: Pressure/Time profiles for the template protocol. A. Step 1 lysate loading, B. Step 3 wash with LSDNA/ethanol, C. Step 5 was with 75% isopropanol, D. Step 6 column drying after washes, E, Step 10 first elution from column, F, Step 14 second elution from the column. Time is measured in seconds. 61
Figure 2.2: Images of the Tecan A200 set up to run the Fire Monkey HMW DNA preparations. Clockwise from top left the images depict side view of A200 with reagent bottle stack with piping into the A200, front view of A200 with compressor and waste container under the desk, 96-well collection plate in bracket, A200 running a flash operation, and the reagent bottle stack. 62
Figure 2.3: Pressure and time settings for Step 1 (lysate loading phase). A, E. coli SR version 2 protocol for the Tecan A200. B, Foundational protocol..... 65
Figure 2.4: Boxplot illustrating the concentration of DNA recovered from the E. coli Fire Monkey preparations Elution 1 and Elution 2, measured in ng/µL..... 66
Figure 2.5: Pressure and time setting for Step 1 (lysate loading phase). A, E. coli SR version 2 protocol for the Tecan A200. B, Salmonella protocol. 69
Figure 2.6: Boxplot illustrating the concentration of DNA recovered from 230 Salmonella Fire Monkey preparations Elution 1 and Elution 2, measured in ng/µL. 70
Figure 2.7: TapeStation electrophoresis traces for Fire Monkey extracted Salmonella HMW DNA. A-C are individual DNA extractions run on a Genomic DNA ScreenTape, the blue traces represent elution 1 and the orange traces represent elution 2. D Genome DNA ScreenTape shows gel images of the same sample from elution 1 and 2. 71
Figure 2.8: Femto Pulse capillary electrophoresis of Fire Monkey Salmonella HMW DNA extractions using the Tecan A200. Images are paired left and right, left is elution 1 and right is elution 2..... 72
Figure 2.9: TapeStation Genome DNA ScreenTape gel image showing aliquots 1-6 from the enzymatic digestion experiment. Using elution 1 for the Fire Monkey samples 2-5. 78
Figure 2.10: Electropherogram from TapeStation Genome ScreenTape for aliquots 1-5 from the enzymatic digestion experiment. Using elution 1 for the Fire Monkey samples 2-5. The dark blue colour represents aliquot 1 (Max-RSC), orange represents aliquot 2, green represents aliquot 3, red represents aliquot 4, and light blue represents aliquot 5. 79
Figure 2.11: TapeStation GenomeTape gel showing DNA from additional washing step testing. 1 = Fire Monkey run with no host depletion and no washing, 2 = Fire Monkey run with host depletion and no washing, 3 = Fire Monkey run with host depletion with one PBS wash, 4 = Fire Monkey run with host depletion with three PBS washes, 5 = Fire Monkey run with host depletion with one warm water wash, and 6 = Fire Monkey run with host depletion with three warm water washes. 81
Figure 2.12: TapeStation GenomeTape electropherograms for samples from additional washing step testing. A depicts the PBS washes (1x = red trace, 3x – green trace) versus control host depleted without washes (blue). B depicts the warm water washes (1x = orange trace, 3x – aqua trace) versus control host depleted without washes (blue). .. 82
Figure 2.13: TapeStation GenomeTape gel showing DNA from additional washing step including alcohols. 1 = no wash Fire Monkey, 2 = 3x warm water washes, 3 = ethanol washes, and 4 = isopropanol washes. 84
Figure 2.14: Agilent TapeStation and Femto Pulse traces of a Max-RSC stool preparation and a 3x warm water washed Fire Monkey stool DNA preparation. Trace 1 is a TapeStation trace of the Max-RSC preparation, trace 2 is a Femto Pulse trace of the

same Max-RSC preparation. Trace 3 is a TapeStation trace of the Fire Monkey preparation, trace 4 in a Femto Pulse trace of the same Fire Monkey preparation..... 85
Figure 2.15: TapeStation GenomeTape gel showing DNA from addition of Neutrase and increasing reagent volume testing. 1 & 2 = Fire Monkey protocol without washing, 3 & 4 FM-W protocol with the Fire Monkey protocol steps carried out at 3x volume, and 5 & 6 Neutrase treated samples..... 87
Figure 2.16: The blue traces represent the Fire Monkey Protocol with no wash, the orange represents the FM-W protocol with the Fire Monkey protocol steps carried out at 3x volume, and the green represents the Neutrase treated sample. This electropherogram is showing aliquots 1, 3, and 6..... 88
Figure 2.17: Black and blue lines represent the Fire Monkey protocol without washes (Aliquots 1 and 2) and the red and orange lines represent the FM-W protocol with the Fire Monkey protocol steps carried out at 3x volume (Aliquots 3 and 4). 89
Figure 2.18: TapeStation GenomeTape on DNA from Max-RSC versus FM-W-3x testing. DNA samples are marked on gel lane. 90
Figure 2.19: Changes made to Tecan A200 pressure profiles for the Stool A200 protocol. A = Step 1 lysate load phase stool protocol, B = Step 1 lysate load phase *E. coli* protocol, C = Step 3 first filter column wash stool protocol, and D = Step 3 first filter column wash *E. coli* protocol. 96
Figure 3.1: Clustal Omega multiple sequence alignments of *napA* from LR-Pilon hybrid assembly, Illumina assembly (short-read only), and ONT assembly (long-read only). Highlighted in blue boxes are the regions within *napA* where putative SNPs were predicted. * = matching base call in all sequences, absent of * = discord between base calls in base position. 125
Figure 3.2: Clustal Omega multiple sequence alignments of *manC1* from LR-Pilon hybrid assembly, Illumina assembly (short-read only), and ONT assembly (long-read only). Highlighted in the blue box is the locations within *manC1* where the putative SNP was predicted. * = matching base call in all sequences, absent of * = discord between base calls in base position. 126
Figure 3.3: Clustal Omega multiple sequence alignments of *ydiN* from LR-Pilon hybrid assembly, Illumina assembly (short-read only), and ONT assembly (long-read only). Highlighted in blue box is the location in *ydiN* where the putative SNP was predicted. * = matching base call in all sequences, absent of * = discord between base calls in base position..... 126
Figure 3.4: Clustal Omega multiple sequence alignments of dnaJ from LR-Pilon hybrid assembly, Illumina assembly (short-read only), and ONT assembly (long-read only). * = matching base call in all sequences, absent of * = discord between base calls in base position..... 127
Figure 3.5: Clustal Omega multiple sequence alignments of tldD from LR-Pilon hybrid assembly, Illumina assembly (short-read only), and ONT assembly (long-read only). * = matching base call in all sequences, absent of * = discord between base calls in base calls in base position. 127
Figure 3.6: Clustal Omega multiple sequence alignments of rcnA from LR-Pilon hybrid assembly, Illumina assembly (short-read only), and ONT assembly (long-read only). * = matching base call in all sequences, absent of * = discord between base calls in base position..... 128
Figure 3.7: Loss of AMR-carrying transposable element in the *S.* Typhimurium genome. Clinker schematic where isolate 3 represents the consensus sequence found in 19/20 isolates from 22EPA044NSA. Flanked by insertion sequences (in orange) several genes including four AMR genes (in red, dark blue, yellow and light blue) were absent in isolate 10. Genes are represented by arrows indicated directionality, with matching

colours indicating identical gene sequence. Homology between the two isolates is represented as black bars, regions without black bars linking them are absent in isolate 10. 139
Figure 3.8: Variation between 20EPA002NSA Isolates. Core genome maximum likelihood tree for twenty *S.* Java ST 43 isolates. Tree overlaid with the SNPs responsible for each branch. Key for SNP type: black = non-synonymous, blue = synonymous, and red = STOP gained. 141
Figure 3.9: Variation between 20EPA011NSA Isolates. Core genome maximum likelihood tree for nineteen *S.* Java ST 149 isolates. Tree overlaid with the SNPs responsible for each branch. Key for SNP type: black = non-synonymous and blue = synonymous. 142
Figure 3.10: Stacked bar chart displaying the count of Serotype (ST) by collection year 146
Figure 3.11: GrapeTrees showing UKHSA United Kingdom origin *S.* Java isolates. Left: An Achtman 7 Gene MLST GrapeTree with the key displaying ST based on the Achtman 7 Gene MLST scheme. Right: A cgMLST GrapeTree with the key displaying ST based on the Achtman 7 Gene MLST scheme. ST43 is shown in dark blue, and ST149 in dark green. The scale bar represents a distance in alleles. 147
Figure 3.12: GrapeTree showing cgMLST for UKHSA United Kingdom origin *S.* Java ST149 isolates supplemented with 20EPA011NSA_1 and 20EPA011NSA_11. The key is at the HC10 levels. The 20EPA011NSA_1 containing cluster is marked with a green arrow, the 20EPA011NSA_11 containing cluster is marked with a blue arrow, and the 20EPA0011NSA_13 and _19 cluster is marked with a red arrow. The scale bar represents a distance in alleles. 149
Figure 4.1: Flowchart of experimental design 166
Figure 4.2: Depth per 10M reads: F0 versus G at timepoints 1, 3, and 9 months. 188
Figure 4.3: $\log_{10}$ depth normalisation: F0 versus G at timepoints 1, 3, and 9 months. 188
Figure 4.4: Depth per 10M reads: F0 versus R at timepoints 1, 3, and 9 months 189
Figure 4.5: $\log_{10}$ depth normalisation: F0 versus R at timepoints 1, 3, and 9 months. 189
Figure 4.6: Depth per 10M reads: F0 versus Z at timepoints 1, 3, and 9 months..... 190
Figure 4.7: $\log_{10}$ depth normalisation: F0 versus Z at timepoints 1, 3, and 9 months. 190
Figure 4.8: Breadth per 10M reads: F0 versus G at timepoints 1, 3, and 9 months. ... 191
Figure 4.9: $\log_{10}$ breadth normalisation: F0 versus G at timepoints 1, 3, and 9 months. 191
Figure 4.10: Breadth per 10M reads: F0 versus R at timepoints 1, 3, and 9 months... 192
Figure 4.11: $\log_{10}$ breadth normalisation: F0 versus R at timepoints 1, 3, and 9 months. 192
Figure 4.12: Breadth per 10M reads: F0 versus Z at timepoints 1, 3, and 9 months. .. 193
Figure 4.13: $\log_{10}$ breadth normalisation: F0 versus Z at timepoints 1, 3, and 9 months. 193
Figure 4.14: Genome fraction per 10M reads: F0 versus G at timepoints 1, 3, and 9 months. 194
Figure 4.15: $\log_{10}$ genome fraction normalisation: F0 versus G at timepoints 1, 3, and 9 months. 195
Figure 4.16: Genome fraction per 10M reads: F0 versus R at timepoints 1, 3, and 9 months. 195
Figure 4.17: $\log_{10}$ genome fraction normalisation: F0 versus R at timepoints 1, 3, and 9 months. 195
Figure 4.18: Genome fraction per 10M reads: F0 versus Z at timepoints 1, 3, and 9 months. 196

Figure 4.19: log10 genome fraction normalisation: F0 versus Z at timepoints 1, 3, and 9 months. 196
Figure 4.20: Boxplot for mean Cp *Campylobacter* (*cadF*) DNA qPCR assay separated by condition and for each condition separated by recovery of full ST score (=7) and incomplete ST score (<7), data shown for G, R, Z includes all storage timepoints (1, 3 and 9 months). F is a single timepoint, before storage (F0)..... 200
Figure 4.21: Boxplot for mean Cp Human DNA qPCR assay separated by condition and for each condition separated by recovery of full ST score (=7) and incomplete ST score (<7), data shown for G, R, Z includes all storage timepoints (1, 3 and 9 months). F is a single timepoint, before storage (F0)..... 201
Figure 4.22: Campylobacter MDG N50 by preservation condition and time point..... 202
Figure 4.23: Distribution of N50 by ST score, including a limit of *Campylobacter* detection line. 203
Figure 4.24: Proportion of reads removed by in-silico human read removal, a proxy for failed host depletion and human read content in the stool sample. 204
Figure 4.25: CheckM completeness of MD-*Campylobacter* genomes from stool 124. Values are percentages standardised by “reads in” and reported per 10 million reads. 205
Figure 4.26: CheckM completeness of MD-*Campylobacter* genomes from stool 132. Values are percentages standardised by “reads in” and reported per 10 million reads. 206
Figure 4.27: CheckM completeness of MD-*Campylobacter* genomes from stool 135 replicate 1. Values are percentages standardised by “reads in” and reported per 10 million reads. 206
Figure 4.28: CheckM completeness of MD-*Campylobacter* genomes from stool 135 replicate 2. Values are percentages standardised by “reads in” and reported per 10 million reads. 206
Figure 4.29: CheckM completeness of MD-*Campylobacter* genomes from stool 136. Values are percentages standardised by “reads in” and reported per 10 million reads 207
Figure 4.30: CheckM completeness of MD-*Campylobacter* genomes from stool 141. Values are percentages standardised by “reads in” and reported per 10 million reads. 207
Figure 4.31: CheckM completeness of MD-*Campylobacter* genomes from stool 143. Values are percentages standardised by “reads in” and reported per 10 million reads. 208
Figure 4.32: CheckM completeness of MD-*Campylobacter* genomes from stool 146. Values are percentages standardised by “reads in” and reported per 10 million reads. 208
Figure 4.33: CheckM completeness of MD-*Campylobacter* genomes from stool 147. Values are percentages standardised by “reads in” and reported per 10 million reads. 209
Figure 5.1: GenomeTape TapeStation trace of the Nanopore library for 164fm..... 228
Figure 5.2: GenomeTape TapeStation trace of the Nanopore library for 164max. 228
Figure 5.3: GenomeTape TapeStation trace of the Nanopore library for 165fm..... 229
Figure 5.4: GenomeTape TapeStation trace of the Nanopore library for 165max. 229
Figure 5.5: Location of the *tet*(*O*) gene in the genome of isolate 165-3. 244
Figure 5.6: Relative abundance of the Top 15 Genera for stool 164 for Fire Monkey and Maxwell preps sequenced by MinION (Long) and Illumina Pair-end 150bp (Short). .. 249

Figure 5.7: Relative abundance of the Top 15 Genera for stool 165 for Fire Monkey and Maxwell preps sequenced by Oxford Nanopore MinION (Long) and Illumina Pair-end 150bp (Short) before removal of *Homo* genus. 251
Figure 5.8: Relative abundance of the Top 15 Genera for stool 165 for Fire Monkey and Maxwell preps sequenced by Oxford Nanopore MinION (Long) and Illumina Pair-end 150bp (Short) after removal of *Homo* genus 252

List of Abbreviations

AMR – Antimicrobial Resistance

AST – Antimicrobial Susceptibility Testing

bp – base pair

BSA – Brilliance Salmonella Agar

CARD – Comprehensive Antibiotic Resistance Database

CC – Clonal Complex

CDC – Centers for Disease Control and Prevention

CFU - Colony-forming unit

cgMLST – Core Genome Multi-Locus Sequence Typing

DNA – Deoxyribonucleic Acid

ECDC – European Centre for Disease Prevention and Control

EPA – Enteric Pathogen Laboratory (contextual from stool sample IDs)

ESBL – Extended-Spectrum Beta-Lactamase

F0 – Baseline (timepoint before storage)

FWD-Net – Food and Waterborne Diseases and Zoonoses Network

G – Glycerol storage condition (Brucella broth + 17.5% glycerol)

GBRU – Gastrointestinal Bacteria Reference Unit

GI – Gastrointestinal

GS – Genome Structure

HC - Hierarchical clustering

kb – Kilobase

MICs – Minimum Inhibitory Concentrations

MLST – Multi-Locus Sequence Typing

MLVA – Multiple-Locus Variable Number Tandem Repeat Analysis

NCBI – National Center for Biotechnology Information

NGS – Next-Generation Sequencing

NTS – Non-typhoidal *Salmonella enterica*

ONT – Oxford Nanopore Technologies

PacBio – Pacific Biosciences

PFGE – Pulsed-Field Gel Electrophoresis

PGAP – Prokaryotic Genome Annotation Pipeline

QC - Quality Control

qPCR – Quantitative Polymerase Chain Reaction

R – Raw stool storage condition (no preservative)

SNP – Single-Nucleotide Polymorphism

SPRI – Solid Phase Reversible Immobilization

ST – Sequence Type

STEC – Shiga Toxin-Producing *Escherichia coli*

UKHSA – UK Health Security Agency

WGS – Whole Genome Sequencing

XLD – Xylose Lysine Deoxycholate Agar

Z – Zymo DNA/RNA Shield storage condition

1 Introduction

Pathogen genomics has rapidly transformed public health microbiology, enabling real-time detection and characterisation of infectious disease threats. By decoding the genetic blueprint of pathogens, genomics allows public health professionals to track transmission pathways, detect emerging variants, and understand the spread of antimicrobial resistance (AMR). These advances underpin modern epidemiological investigations and strengthen public health responses to outbreaks, especially in the context of food-borne illnesses.

1.1 The Role of Genomics in Outbreaks and Antimicrobial Resistance

1.1.1 Outbreak Investigation and Surveillance

Pathogen genomics offers a high-resolution lens through which outbreaks can be investigated. Whole genome sequencing (WGS) surpasses traditional subtyping methods such as pulsed-field gel electrophoresis (PFGE) or multilocus sequence typing (MLST) by offering single-nucleotide polymorphism (SNP)-level resolution that distinguishes closely related strains (Allard, 2016). SNP-level precision is especially valuable for detecting widespread, multi-jurisdictional outbreaks where epidemiological links are not immediately apparent (Popa & Popa, 2021).

Several real-world examples highlight this impact. During the 2011 European outbreak of Shiga toxin-producing *Escherichia coli* (STEC) O104:H4 the use of WGS enabled the identification of contaminated fenugreek sprouts/seeds as the source. SNP-level resolution was crucial for informing targeted control measures, understanding transmission routes, and identifying the difference in diversity between the German and French outbreak samples (Beutin & Martin, 2012; Grad et al., 2012). In the UK, the use of WGS has been instrumental for managing *Salmonella enterica* outbreaks. In May 2015 WGS was used to identify and investigate a *Salmonella* Enteritidis outbreak linked to contaminated chicken eggs. Genomic analysis, combined with food-chain investigation pinpointed the source and supported rapid intervention (Inns et al., 2017). In another example, the use of WGS played a key role for the U.S. Centers for

Disease Control and Prevention (CDC) when it was used to resolve a *Salmonella enterica* outbreak linked to contaminated cucumbers in 2015. Employing WGS made it possible to link over 900 cases of *Salmonella* across 40 states by comparing SNPs between isolates. This level of detail revealed connections that traditional methods had missed (Kozyreva et al., 2016). More recently, WGS was central to resolving a large international outbreak of monophasic *Salmonella* Typhimurium linked to chocolate products in 2022. The integration of genomic, epidemiological, and food-chain data across multiple countries enabled rapid source tracing to a single manufacturing plant and guided a global product recall (Laisnez et al., 2025).

1.1.2 Antimicrobial Resistance Monitoring

Culture is the gold standard for establishing an infectious agent's AMR profile, and typical growth-based antimicrobial susceptibility testing (AST) requires many cultivation steps. These steps typically include growing on agar plates to obtain single colony-forming units (CFUs), enriching to increase the bacterial load, and testing different antibiotic doses in liquid or solid medium (Vasala et al., 2020). This process can be resource-intensive, which could cause delays in getting results.

A modern solution to this problem involves the use of genomics. WGS can be used to obtain sequence-based AMR predictions in a culture-dependent manner, while metagenomics enables culture-independent AMR prediction directly from sequence obtained from complex samples. Sequence based AMR detection tools such as ResFinder, abritAMR, ARIBA, and the Comprehensive Antibiotic Resistance Database (CARD) make it easy to quickly find AMR determinants (Zankari et al., 2012). These tools support AMR surveillance from cultured bacteria and directly from diverse sample types including stool, food, and environmental samples (Anjum, 2015; Dziegiel et al., 2024; Noyes et al., 2016). This helps find new resistance risks earlier and improve surveillance and response to outbreaks. Although there is usually a strong link between genotype and phenotype, a key concern with the genomic based approach is that resistance profiles may not be accurate.

A comprehensive investigation conducted by the UK Health Security Agency (UKHSA) using 3,491 non-typhoidal *Salmonella* enterica (NTS) isolates demonstrated exceptional overall concordance, with 0.17% of phenotypic and genotypic

isolate/antimicrobial combinations exhibiting discordance. Some disparities were found, particularly with streptomycin, highlighting the limitations of sequence-based inference for certain antibiotics (Neuert et al., 2018). A 99.74% concordance rate between sequence-based predictions and phenotypic AST results was reported in a Danish study that examined 200 isolates from pigs that focused on four different bacterial species. The majority of mismatches in that study were associated with spectinomycin resistance in *E. coli* (Zankari et al., 2013). Collectively, these findings highlight WGS's potential as a quick and accurate AMR surveillance tool. WGS should however, complement rather than replace phenotypic AST, due to occasional discordances and the need for clinical clarity, especially when it comes to directing empirical treatment decisions in clinical settings.

The detection of plasmid-mediated colistin resistance gene *mcr-1* in livestock and clinical isolates across multiple countries underscored the urgency of One Health surveillance strategies (Bastidas-Caldes et al., 2022; Daza-Cardona et al., 2022; Noyes et al., 2016). The One Health approach recognises that human health, animal health, and environmental health are all interconnected. Tackling AMR requires coordinated efforts across sectors. The finding of *mcr-1* serves as an example of how resistance genes can arise in agricultural environments, most likely as a result of livestock antibiotic use. Then, it can spread to people by environmental channels, the food chain, or direct contact. The One Health framework brings together researchers from veterinary, clinical, and environmental microbiology. By integrating these fields it allows us to monitor transmission routes more effectively and respond in ways that help slow the global spread of AMR (Destoumieux-Garzón et al., 2018). Use of WGS can reveal whether resistance genes are located on chromosomes or plasmids, in addition to identifying them (Berbers et al., 2020). This is significant because AMR spreads more quickly across species and settings due to the ease with which plasmid-borne genes can be transferred between bacteria.

1.1.3 Overview of Foodborne Pathogens: *Salmonella* and *Campylobacter*

1.1.3.1 *Salmonella enterica*

Salmonella enterica is a leading cause of foodborne illness globally. Common transmission vectors are poultry, eggs, meat, water, and contact with infected animals

and people (Popa & Popa, 2021). With thousands of distinct serovars it is epidemiology complex varying by geography and food production practices (Achtman et al., 2012). Historically, serotyping provided a framework for identification. Today, WGS enables finer discrimination within and between serovars (Chattaway et al., 2023).

Taxonomically, the genus is composed of two recognised species *Salmonella enterica* and *Salmonella bongori,* which diverged from a common ancestor tens of millions of years ago (Wang et al., 2019). *S. enterica* is divided into multiple subspecies (historically six major subspecies designated I, II, IIIa, IIIb, IV, and VI) encompassing over 2,500 known serovars (Lamas et al., 2018). Subspecies *enterica* (I) includes more than 1,500 serovars and accounts for >99% of human *Salmonella* infections. In contrast, the other *S. enterica* subspecies (II, IIIa, IIIb, IV, VI), along with *S. bongori*, are primarily associated with cold-blooded animals or environmental niches and only rarely cause disease in humans. Non-*S. enterica* lineages usually only infect humans as opportunistic diseases in immunocompromised patients and lack specific pathogenicity factors (Lamas et al., 2018).

The evolutionary links within *Salmonella* have been elucidated by advances in phylogenomic analysis, which have shown several profoundly branching lineages. Within *S. enterica*, WGS studies confirm that each subspecies represents a genetically distinct clade (Pearce et al., 2021). Pearce *et al.* (2021) analysed a large collection of clinical isolates uncovered several previously unrecognised lineages now proposed as new subspecies, namely *S. enterica* subsp. *londinensis* (VII), *brasiliensis* (VIII), *hibernicus* (IX), *essexiensis* (X), and a newly identified subsp. *reptilium* (XI). This study also reported that the conventional *S. enterica* subsp. *arizonae* (IIIa) is highly divergent from the other *enterica* subspecies; it clusters apart and may warrant classification as a separate species, *S. arizonae*. In comparison to more traditional biochemical techniques, this refined phylogeny demonstrates the higher accuracy of genomic approaches for resolving *Salmonella* taxonomy.

1.1.3.2 *Campylobacter*

Campylobacter jejuni is the most common bacterial cause of gastroenteritis in many high-income countries and is strongly associated with poultry consumption (Facciolà et al., 2017). Unlike *Salmonella*, *C. jejuni* exhibits high levels of genome plasticity due

to phase variation, recombination, and hypervariable loci (Cody et al., 2013). AMR in *C. jejuni* has become a concern, particularly resistance to fluoroquinolones and macrolides driven by point mutations such as *gyrA* T86I and A2074G/A2075G in the *23S rRNA* gene (Bukari et al., 2025). Widespread use of WGS has accelerated the detection of these resistance mechanisms supporting their inclusion in routine surveillance workflows (Zankari et al., 2017).

Taxonomically, the genus is composed of 33 species that cluster into five principal clades, conventionally named after representative species: the *C. jejuni* group, *C. lari* group, *C. concisus* group, *C. ureolyticus* group, and *C. fetus* group (Costa & Iraola, 2019; Wu et al., 2024). The clinical relevance of this genus is underscored by all lineages containing pathogenic species. The *C. jejuni* group contains the major zoonotic *Campylobacter* of humans, *C. jejuni* and its close relative *C. coli*. *C. jejuni* group species are thermotolerant and prevalent in poultry and other warm-blooded animals. The *C. fetus* group includes *C. fetus* subsp. *fetus* and *C. fetus* subsp. *venerealis*, mostly recognised as veterinary pathogens causing infertility and abortions in cattle and sheep, and occasionally invasive infections in humans. The remaining groups (e.g. *C. concisus*, *C. lari*, *C. ureolyticus*) contain various emerging or niche-adapted species (such as oral bacterium *C. concisus*, avian-associated *C. lari*, and gastrointestinal *C. ureolyticus*), some of which are increasingly being implicated in human disease (Costa & Iraola, 2019).

Campylobacter populations exhibit high levels of genetic diversity and plasticity (Woodcock et al., 2017). This occurs within single geographic regions and within host population (Sheppard et al., 2009). *C. jejuni* isolates are often highly heterogeneous with numerous distinct lineages co-circulating (Cody et al., 2013). There is little geographic or clonal structure in the population with isolates from distant locations often intermingling on the phylogenetic tree (Sheppard et al., 2013). This indicates frequent gene flow across populations. The observed extensive diversity is driven in part by *Campylobacter's* propensity for horizontal gene transfer and recombination (Golz & Stingl, 2021). Many of the most frequently recombined genes are involved in surface structures and adaptation, such as genes for heptose biosynthesis (a component of lipooligosaccharide), host colonisation factors, and stress response suggesting strong selection on antigenic and survival traits (Park et al., 2020). High

levels of intra-species recombination result in a non-clonal population structure. Genetic exchange among strains (even between species like *C. jejuni* and *C. coli*) decouples genotype from lineage such that traditional typing markers like serotype often do not correspond to a strictly vertical phylogeny (Barker et al., 2020).

1.1.4 Brief History of Sequencing Technologies in Epidemiology

In 1977, the first practical method for DNA sequencing was developed by Frederick Sanger (Sanger et al., 1977). Sanger sequencing laid the foundation for molecular epidemiology enabling the development of MLST schemes that became standard for bacterial typing in the early 2000s. The adoption of MLST also led to the creation of global databases, including PubMLST and EnteroBase (Page et al., 2017; Pérez-Losada et al., 2013). These platforms grew quickly as sequencing data accumulated. Sanger sequencing was a remarkable technological advancement, but was labour-intensive and low throughput, limiting its utility during large-scale outbreaks (Chiang & Palmore, 2022).

WGS is now central to surveillance networks such as UKHSA Gastrointestinal Infections Network, European Centre for Disease Prevention and Control (ECDC) Food and Waterborne Diseases and Zoonoses Network (FWD-Net) and U.S. CDC's PulseNet, which routinely apply WGS to foodborne pathogens for outbreak detection and response (Brown et al., 2019; Chattaway et al., 2023; Revez et al., 2017). More recently, third-generation technologies such as Oxford Nanopore and Pacific Biosciences have enabled real-time sequencing and improved the resolution of repetitive or structurally complex genomic regions (Espinosa et al., 2024). Long-read sequencing is particularly helpful for defining mobile elements such as phages, transposons, and plasmids and for assembling complete genomes (Huisman et al., 2022; Waters et al., 2025). These techniques are increasingly being used in hybrid assemblies, which combine long-read and short-read data to produce complete, high-quality genome reconstructions called hybrid genomes. Antimicrobial resistance genes and their genetic context, such as plasmid vs chromosomal placement, can be more precisely resolved thanks to hybrid genomes, which combine the base-level accuracy of short reads with the structural completeness of long reads. National and international databases like EnteroBase, which integrate MLST, core genome, and SNP-based phylogenies, and platforms like Nextstrain, which visualise pathogen

evolution in near real-time, exemplify how DNA sequencing underpins modern genomic epidemiology (Alikhan et al., 2018; Hadfield et al., 2018). These tools enable proactive monitoring of outbreaks and the spread of AMR, supporting evidence-based interventions and global health preparedness.

1.2 Short-Read Sequencing in Public Health Applications

Short-read WGS has become a cornerstone of modern public health microbiology. In this approach, DNA from purified single bacterial colonies is fragmented and sequenced in many short pieces (typically 150–300 base pairs), which are then reconstructed *in-silico* (Goodwin et al., 2016). Thanks to next-generation sequencing (NGS) technologies like Illumina, public health organisations worldwide have rapidly adopted short-read WGS for routine pathogen surveillance and outbreak response. For example, UKHSA began sequencing all *Salmonella* isolates referred to its laboratories in 2014, revolutionising reference microbiology and surveillance practices (Chattaway et al., 2019a). Likewise, the U.S. CDC's PulseNet network transitioned in 2019 from traditional subtyping (PFGE) to WGS as the primary method for all bacterial foodborne pathogens (Ribot et al., 2019). The greater resolution and effectiveness that short-read sequencing provides for tracking infectious diseases is what is driving this broad adoption.

1.2.1 Advantages of Short-Read Sequencing

Short-read sequencing is highly accurate at reading DNA with Illumina's sequencing-by-synthesis chemistry achieving very low error rates of ~0.1–1% per base (Zhang et al., 2020). Modern Illumina instruments (e.g. NovaSeq, NextSeq, HiSeq) report ≥85% Q30 corresponding to an error probability of 1 in 1,000, or 0.1% error rate (Polonis et al., 2025). Illumina short-read platforms are considered cost-effective and high-throughput, allowing hundreds of bacterial genomes to be sequenced in a single run at relatively low cost per sample (Struelens et al., 2024). In comparison to older sequencing methods, Illumina platforms greatly reduce sequencing time by sequencing many DNA fragments in parallel. Being able to multiplex samples is particularly valuable for public health labs. Hundreds of isolates can be sequenced in a single run. This efficiency makes WGS practical for real-time surveillance and outbreak detection (Gilchrist et al., 2015). Additionally, short-read sequencing benefits

from known techniques and verified bioinformatics pipelines in labs with established infrastructure, which facilitates its integration into regular processes.

1.2.2 Limitations of Short-Read Sequencing

In many low- and middle-income countries, the high start-up costs and limitations in informatics capacity remain barriers to adoption (Sekyere & Reta, 2020; WHO, 2022). Despite its strengths, short-read sequencing has important limitations, mostly stemming from the short length of reads. Because each read is only a few hundred bases long, it can be challenging to assemble complete genomes or map reads uniquely in repetitive regions (Treangen & Salzberg, 2012). Draft genome assemblies from short reads are frequently fragmented with gaps, because repeating DNA sequences or mobile elements are longer than an individual read and hence cannot be resolved (Neal-McKinney et al., 2021) . For example, genes in highly repetitive regions or paralogous gene families may not map confidently and could be missed. In bacteria, this means plasmids or other mobile genetic elements carrying antimicrobial resistance genes might not be correctly linked to their host genome using short reads alone (Juraschek et al., 2021). Short-read WGS is also less effective for detecting large structural variants or gene arrangements compared to long-read approaches (Sedlazeck et al., 2018). Moreover, in routine practice WGS requires a pure culture of the organism; contamination with other DNA can confound the analysis, which remains a logistical limitation, especially for culture-free diagnostic samples. WGS-based investigations rely on databases of known genetic markers (for serotype, pathogenicity, and resistance), therefore truly unique mutations or genes may be missed (Chattaway et al., 2019a). Understanding these limitations is crucial as public health labs interpret short-read sequencing data and, when necessary, employ complementary methods to achieve complete genomic insight. Lastly, implementing short-read WGS in routine public health practice presents several challenges. These include the need for substantial infrastructure and specialised training, limited bioinformatics capacity, issues with data storage and secure sharing, and the lack of standardised validation and regulatory frameworks across regions (Black et al., 2020; Libuit et al., 2023).

1.2.3 Applications in Public Health Genomics

1.2.3.1 Bacterial Typing and Surveillance

In countries where short-read WGS has been implemented, it has largely replaced many traditional bacterial typing methods for surveillance. Using genome data, laboratories can identify the species, serotype, and strain lineage of an isolate in a single process, instead of performing separate biochemical tests and serological typing. Prior to WGS, *Salmonella* reference labs required multiple laborious methods including biochemical tests, serotyping, phage typing, PFGE, and multiple-locus variable number tandem repeat analysis (MLVA) to characterise isolates (Chattaway et al., 2019a). Now, a single WGS run can provide the same information with higher resolution. Genome-based typing (for instance, assigning sequence types by MLST or comparing core genomes) offers far greater discrimination between strains than older techniques, which is especially important for detecting clusters of related cases (Ribot et al., 2019). The discriminatory power of WGS has enabled surveillance programs to define genetic subtypes down to the level of single nucleotide differences. As a result, public health databases have expanded with genomic profiles: for example, the EnteroBase project has assembled over 300,000 *Salmonella* genomes from Illumina short reads, underpinning global strain tracking efforts (Zhou et al., 2020). In routine practice, agencies like PulseNet and UKHSA report strain information using WGS-based nomenclature (such as MLST clonal complexes or core genome profiles) as part of weekly surveillance. This genomic technique simplifies operations by extracting several reference properties from sequence data (species, serotype, virulence factors) in one phase (Ribot et al., 2019). Overall, short-read sequencing has made bacterial typing more precise and has unified surveillance data.

However, despite its many advantages, short-read WGS is not without limitations. As previously mentioned short reads are often unable to resolve highly repetitive genomic regions or fully characterise mobile genetic elements such as plasmids or transposons, structures that can carry virulence or antimicrobial resistance genes critical for surveillance (Arredondo-Alonso et al., 2017; Berbers et al., 2020; Luan et al., 2024). Moreover, the accuracy of WGS-based typing is dependent on high-quality sequencing data, robust assembly pipelines, and well-maintained reference databases. Differences in bioinformatics methods between laboratories can lead to

inconsistent results or complicate inter-laboratory comparisons (Mixao et al., 2025). The implementation of WGS also requires significant investment in sequencing infrastructure, data storage, and bioinformatics expertise resources that are often out of reach for many low- and middle-income countries (Sekyere & Reta, 2020). Furthermore, despite its diagnostic potential, WGS data interpretation still relies on curated databases and expert review, and it lacks standardisation in some areas (e.g., serotype calling or resistance prediction) across regions and platforms (Cooper et al., 2020; Sherry et al., 2023; Strepis et al., 2025).

1.2.3.2 Outbreak Investigations

Perhaps the most celebrated application of short-read WGS in public health is the investigation of outbreaks. While WGS offers exceptional resolution for identifying clusters of related cases, genomic data alone are not sufficient to define outbreaks. Epidemiological information such as patient histories, exposures, and temporal-spatial patterns remains essential to contextualise genetic findings and establish plausible transmission routes. By comparing whole-genome sequences, investigators can determine how closely related different isolates are, which helps pinpoint the source and scope of an outbreak. Short-read WGS can reveal differences of just a few SNPs between isolates, a level of resolution that surpasses traditional subtyping methods. This has transformed outbreak detection: clusters of cases that would previously go unrecognised can now be identified through genomic similarity.

Salmonella has been a trailblazer for WGS integration into public health. In England and Wales, the UKHSA Gastrointestinal Bacteria Reference Unit (GBRU) shifted to a WGS-based workflow for *Salmonella* surveillance starting in 2014–2015, processing roughly 8,000–10,000 isolates per year (Chattaway et al., 2019a). This replaced a decades old regime of serotyping and phage typing with a faster, more discriminatory genetic approach. By sequencing every isolate, UKHSA could characterise strains by their sequence type (ST) and core genome, uncovering relationships that traditional serotyping might mask. For example, what was once reported simply as "*S.* Enteritidis" is now recognised as multiple genetically distinct lineages within that serovar (Chattaway et al., 2019a). WGS data has allowed the reference laboratory to infer serotype from sequence (using tools like SeqSero or SISTR) and largely phase out phenotypic serological tests. Within a few years, ~89 % of *Salmonella* isolates were

fully typed by WGS (serovar inferred from genotype), with only ~11 % requiring any traditional methods (usually for novel or mixed-strain cases). The impact on outbreak detection was immediate: WGS provides nearly real-time assessment of clusters. In 2015, the UK was able to detect a nationwide outbreak of *S.* Enteritidis linked to eggs and respond more effectively, thanks to the high resolution of SNP analysis distinguishing the outbreak strain (Inns et al., 2017). An added benefit in *Salmonella* surveillance has been the ability to monitor evolution and introduction of new strains. For example, genomic surveillance noted the first case of extended-spectrum beta-lactamas (ESBL)-producing *S.* Typhi in the UK, enabling rapid public health response to contain its spread (Chattaway et al., 2019a; Nair et al., 2021).

A striking example comes from *Campylobacter*, a pathogen where outbreaks were historically thought to be rare. Denmark's national institute (Statens Serum Institut: SSI) began routine WGS of *Campylobacter* from patients in 2019 and discovered multiple small outbreaks and one unusually large, continuous outbreak, findings that would otherwise have remained unknown without genomics (Joensen et al., 2020; Joensen et al., 2021). This overturned the assumption that most *Campylobacter* infections are sporadic, showing that many infections in fact stem from common sources (in Denmark's case, largely chicken meat).

Similarly, in the UK, WGS-based cluster analysis has enhanced outbreak response. UKHSA's system assigns a "SNP address" to cluster related cases, which has been used to link cases across regions and even internationally in real time (Chattaway et al., 2019a). One investigation in England traced a *Campylobacter* outbreak to raw milk: genome sequencing showed that isolates from patients and farm milk had an identical sequence type ST-7432 (clonal complex 403), confirming the source of infection (Kenyon et al., 2020). For *Salmonella*, WGS has similarly enabled rapid detection of outbreaks that previously might have been missed if strains shared a common serotype but were not identical genetically (de la Gandara, 2023; Inns et al., 2017). The high resolution of short-read WGS allows epidemiologists to distinguish outbreak strains from background cases and to map the spread of pathogens through the food chain or healthcare settings with unprecedented clarity. Ultimately, it is the integration of genomic resolution with classical epidemiological investigation that provides the clearest picture of how outbreaks emerge and spread.

Beyond outbreaks, WGS is being used to study *Campylobacter*'s population structure and source attribution. For example, researchers have employed core genome MLST on large collections of *Campylobacter* genomes to estimate what proportion of human infections come from chickens, cattle, wild birds, and additional sources, improving our understanding of transmission reservoirs (Arning et al., 2021; Thépault et al., 2017; Thystrup et al., 2025).

1.2.3.3 Antimicrobial Resistance Detection

Another important application of short-read sequencing in public health is the detection of AMR. WGS data can be mined for known resistance genes and mutations, which allows prediction of an isolate's antibiotic resistance profile. This genomic technique to AMR detection is faster than AST and can detect resistance pathways even without selective culture. In order to track the emergence of disease resistance, public health labs now regularly check WGS results for a panel of AMR genes. For instance, in the UK, over 17,000 *Salmonella* isolates were sequenced between 2016 and 2018; no phenotypic resistances were missed by WGS screening (though not every genotype is expressed phenotypically), and this provided real-time surveillance of resistance determinants nationwide (Chattaway et al., 2019a). One notable success was the early detection of mcr-1, a plasmid-mediated colistin resistance gene. When mcr-1 was first reported internationally, UKHSA researchers quickly queried their WGS database of ~24,000 enteric bacteria (including *Salmonella, E. coli, Shigella, Campylobacter*) and identified 15 isolates from humans and food carrying this gene (Chattaway et al., 2019a). This demonstrated how short-read WGS archives might quickly be utilised to identify emergent dangers in-silico, without the need to manually test each isolate in the laboratory. There are limitations: purely genotypic prediction may miss novel resistance elements or polygenic traits, and correlations between genotype and drug minimum inhibitory concentrations (MICs) are still being refined (Chattaway et al., 2019a; Kim et al., 2022). Nonetheless, short-read sequencing provides a powerful early warning system for AMR. It allows public health agencies to map resistance genes across bacterial populations and detect worrisome trends (e.g. the rise of ciprofloxacin-resistant *Campylobacter* or multi-drug resistant *Salmonella*).

1.3 High-Throughput Sequencing and Metagenomic Approaches in Pathogen and Community Analysis

The advent of inexpensive, ultra-high-throughput DNA sequencers has transformed WGS of microbes from a costly, specialised endeavour into a routine practice (Gilchrist et al., 2015). Modern benchtop sequencers can rapidly produce gigabases of data per run, making genomic data generation fast and affordable nearly anywhere in the world (Urban et al., 2023). As a result, public health laboratories are increasingly leveraging WGS for surveillance and outbreak investigations, yielding unprecedented resolution in pathogen genotyping. Metagenomics, which applies these sequencing tools to all genetic material in a sample in an untargeted manner, broadens our possibilities by allowing a hypothesis-free search for any pathogen present. This approach allows simultaneous identification of diverse microorganisms (viruses, bacteria, fungi, parasites) with high precision (Ko et al., 2022). Together, high-throughput WGS and metagenomic sequencing provide a comprehensive view of infectious agents and microbial communities that was not attainable with traditional diagnostic methods.

Importantly, these innovations serve a critical need in infectious disease control. GI infections and outbreaks impose significant global morbidity and mortality, particularly among young children (Moore et al., 2015). However, determining the exact cause of an outbreak can be challenging. A significant portion of epidemics have no known cause since traditional diagnostic techniques like culture, antigen testing, or PCR that target certain organisms occasionally fall short of identifying the culprit (Anthony et al., 2024; Franklin et al., 2020; Perrocheau et al., 2023). High-throughput sequencing has emerged as a powerful option; by providing large volumes of sequence data, it can disclose all organisms present in a sample, including new or unexpected infections (Lipkin, 2010; Moore et al., 2015). These data's thoroughness makes it possible to find agents that had not been found before, which significantly improves our capacity to determine the sources of outbreaks.

1.3.1 High-Throughput Sequencing Pipelines and Their Scalability

High-throughput sequencing technologies have made it feasible to sequence pathogen genomes at scale, which is crucial for surveillance and outbreak response.

Throughput and automation in sequencing pipelines mean that laboratories can process hundreds or thousands of isolates in parallel. For example, PulseNet (the U.S. national foodborne disease surveillance network) fully transitioned to WGS as its standard subtyping method in 2019. PulseNet's laboratories collectively sequence on the order of 65,000 bacterial isolates (e.g. *Salmonella, E. coli, Listeria, Shigella*) every year as part of routine monitoring (Kubota et al., 2019). This scalability demonstrates that genomics can be integrated into high-volume public health workflows. The pipeline typically consists of automated DNA extraction, library preparation, fast sequencing, and bioinformatics analysis, which can be finished in 72 hours using express procedures. In optimal settings, sequencing can begin immediately upon sample receipt, bypassing the need for culture and enabling rapid identification of pathogen species, strain, virulence factors, and resistance genes. However, in routine public health practice, logistical factors such as batching, sample transport, and quality control can prolong turnaround times up to 7-11 days (P. Benoit et al., 2024; Huang et al., 2017). When running efficiently WGS workflows can be a drastic improvement over conventional culture-based subtyping, which can take 5-10 days and may fail for fastidious organisms(Forbes et al., 2017; Hilt & Ferrieri, 2022; Tang et al., 2019).

The high resolution of WGS can resolve microbial strains that differ by as little as a single SNP. In practical terms, this means WGS-based subtyping can discern outbreak strains with extraordinary precision, often replacing multiple targeted tests with one sequence-based assay. Access to abundant sequence data has already improved the ability to detect and track outbreaks in real time (Black et al., 2020). As more sequencing data accumulates, it enables creation of large genomic databases against which new isolates are compared. If two patients' bacterial isolates have virtually identical genomes, investigators can quickly recognise them as part of the same cluster (even if they occurred in different regions), prompting an outbreak investigation sooner than was possible with older typing methods. The use of WGS during the COVID-19 pandemic was a significant milestone. Laboratories globally sequenced millions of SARS-CoV-2 genomes, scaling up workflows to unprecedented levels and demonstrating that high-throughput sequencing can inform public health on a global scale (Furuse, 2021; Nicholls et al., 2021). The data generation itself is no longer the bottleneck, sequencing can be done rapidly and cheaply with the attention shifting to

ensuring we have the computational tools and expertise to analyse the flood of genomic data (Black et al., 2020; Gilchrist et al., 2015). In summary, scalable sequencing pipelines now form the backbone of modern pathogen surveillance, offering speed, volume, and resolution that have transformed outbreak detection and investigation.

1.3.2 Metagenomics for Stool Samples and Microbial Communities

Metagenomic study of clinical samples, especially faeces in relation to GI illnesses, is one of the most exciting uses of high-throughput sequencing. Metagenomic next-generation sequencing refers to sequencing all genetic material (microbial and host, DNA and RNA) present in a sample, without needing to isolate or culture specific organisms (Chiu & Miller, 2019). This entails capturing the complete gut microbial community, or microbiome, as well as any pathogens present in a stool sample. The approach is hypothesis-free as it does not require the clinician to decide which pathogen to test for. This is invaluable for diarrhoeal illnesses because symptoms of different GI pathogens overlap, and co-infections can occur (Djeghout et al., 2025; Mai et al., 2025). Metagenomics has the potential to simultaneously detect bacteria, viruses, parasites, or fungi, including rare or unexpected aetiologies, but in practice its effectiveness depends on multiple factors including pathogen abundance, nucleic acid preservation, sequencing depth, DNA extraction bias, and bioinformatic interpretation. As Ko and colleagues noted, a metagenomics-enabled method offers the chance to catch both known and yet to emerge pathogens in a single experiment (Ko et al., 2022).

Applying metagenomics to stool has several key advantages. First, it is culture independent. Many enteric pathogens are difficult or slow to grow in labs, and some routine tests (e.g. for viruses) might be too specific or insensitive (Costantini et al., 2010). By sequencing directly from the sample, metagenomics can reveal organisms that routine diagnostic tests missed. For instance, an analysis of faecal samples from unsolved gastroenteritis outbreaks showed that unbiased metagenomic sequencing could detect the presence of viruses, bacteria, and parasites that had evaded standard diagnostic testing (Moore et al., 2015). In that study, although no completely novel virus was discovered, the sequencing identified known pathogens (such as adenovirus, rotavirus, sapovirus, and a parasite *Dientamoeba fragilis*) that had not

been caught during the original outbreak investigations. This underscores how metagenomics can act as a tool to find missed causes of outbreaks and provide quantitative insights by identifying pathogens and giving a readout of their relative abundance in the sample (Blanco-Míguez et al., 2023). In another example, researchers used metagenomic sequencing on stool from diarrhoea patients and found not only the expected foodborne pathogen but also a secondary pathogen (*Staphylococcus aureus*) present in some cases (Huang et al., 2017). Such co-infections might explain unusual clinical severity or symptoms and would likely have been missed if only a single pathogen test were done.

It should be noted that while the potential of metagenomic sequencing for diagnosis is immense, it is still an emerging technology in practice. One review described diagnostic metagenomics as a "rapidly evolving" tool for culture-independent detection and tracing of foodborne pathogens, with the potential to become a generic platform for identifying most pathogens across many sample types (Andersen & Hoorfar, 2018). However, as of today it remains in an early experimental stage. Challenges such as distinguishing true pathogen sequences from background microbial noise (high abundance commensal bacteria, environmental contaminants, and sequencing artifacts), handling the large volumes of data, and interpreting the clinical significance of every organism detected are areas of ongoing research. Despite these hurdles, the trajectory is clear, metagenomic analysis of stool is moving from research into clinical and public health laboratories, and it is expected to fundamentally improve how we diagnose mysterious gastroenteritis cases (Batool & Galloway-Peña, 2023; Chiu & Miller, 2019; Trivett et al., 2025). By capturing the full picture of the gut microbial community during infection, metagenomics not only finds the needle in the haystack (the pathogen in a complex sample) but also characterises the haystack itself. This could yield new insights into pathogen–microbiome interactions, such as how the composition of gut microbiota might influence transmission or severity of an infection.

1.3.3 Applications in Transmission Dynamics and Outbreak Source Tracking

Beyond identifying the causative agent of an outbreak, high-throughput sequencing data can illuminate how an outbreak spreads and where it originated. Genomic data

serves as a kind of fingerprint for a pathogen strain; by comparing genomes from different cases and sources, epidemiologists can infer relationships and map out transmission networks. WGS is now commonly used to investigate transmission chains in hospitals, communities, and across borders. The resolution of WGS is so high that it can often distinguish whether patients were infected from a common source or from separate introductions. As one review summarised, using WGS in outbreak analysis allows investigators to identify paths of disease transmission within a population and even pinpoint the probable source of the outbreak (Gilchrist et al., 2015).

A recent investigation used a metagenomic microbial source tracking approach to solve a national outbreak of cryptosporidiosis (a parasitic diarrhoeal disease) linked to romaine lettuce. In this 2021 case, over a hundred people were sickened by the parasite *Cryptosporidium parvum*. Scientists sequenced DNA from patients' stool samples as well as from suspect lettuce and other environmental samples, then compared the microbial communities. They found that the genetic signature of microbes on the contaminated lettuce matched that in patients, helping confirm the lettuce as the vehicle of infection (Ahlinder et al., 2022). By examining not just the pathogen's genome but the entire metagenomic, they could infer the contamination likely resulted from sewage water. This work demonstrated how metagenomics may be utilised for forensic tracking of outbreak sources, particularly organisms that are difficult to classify using standard methods. It mirrors the increasing "One Health" concept, which combines microbiology data from humans, food, and the environment to better identify transmission paths.

1.4 Advancing Pathogen Analysis with Long-Read Sequencing

Recent developments in long-read sequencing technology, particularly those created by Pacific Biosciences (PacBio) and Oxford Nanopore technology (ONT), are significantly improving our ability to reconstruct whole bacterial genomes with increased precision and completeness. These platforms produce reads that are many kilobases long, frequently surpassing 10 kb and sometimes surpassing 100 kb. This allows for the resolution of large-scale structural changes and the span of repeated regions that are not achievable with short reads (Scarano et al., 2024; Wick et al., 2017a). Importantly, these technologies have the potential to enable de novo genome

assembly, comprehensive plasmid reconstruction, and extensive tracking of mobile genetic elements (Kwon et al., 2020; Zhao et al., 2023), all of which are critical in the genomic research of foodborne pathogens.

1.4.1 Overcoming the Limitations of Short-Read Sequencing

Short-read sequencing frequently generates fragmented assemblies made up of hundreds of contigs, especially when applied to complicated samples like stool. The genomic context of important virulence and AMR genes is obscured by assembly gaps and misassembles caused by short reads' incapacity to span repetitive regions (Berbers et al., 2023). For instance, the localisation of AMR genes on plasmids versus the chromosome is critical to understanding transmission risk and is frequently unresolved in short-read data alone. Long-read sequencing addresses these limitations by producing contiguous assemblies, frequently near-complete or even fully assembled for well-covered genomes that can span ribosomal operons, insertion sequences, genomic islands, and entire plasmids (Sia et al., 2025; Wick et al., 2023; Zhao et al., 2023). This has direct implications for pathogen typing and outbreak investigations, as complete genomes allow for higher resolution phylogenetics, enhanced serotyping, and the detection of rearrangements or novel elements that may contribute to pathogen fitness or persistence.

1.4.2 Error Correction and Assembly Polishing

Raw ONT reads are prone to systematic errors (often 5–15% error rate) dominated by indels, especially in homopolymers (Luan et al., 2024). Gene annotation may be disrupted by frameshifts in coding areas caused by uncorrected indels (Wick & Holt, 2022). This is particularly crucial for pathogens. *Salmonella* genomes (<4.8-5.0 Mb) contain multiple repetitive pathogenicity islands and plasmids crucial to virulence, while *Campylobacter jejuni* genomes (~1.7 Mb) are short, AT-rich, and carry several repetitions and plasmids that hamper assembly (Neal-McKinney et al., 2021). In *Campylobacter*, simple sequence repeats mediate phase-variable genes, which are loci whose expression stochastically switches on and off through slipped-strand mispairing, creating additional indel hotspots. These factors mean that pathogen assemblies are often fragmented or contain misassembled mobile elements if not polished (Cayrou et al., 2021; Yamamoto et al., 2021).

Hybrid assembly takes advantage of the complimentary qualities of long and short readings. Long ONT reads (tens of kilobases) can cover repetitions, structural variations, and whole plasmids, resulting in continuous assemblies. For example, one study showed that hybrid assemblies of *C. jejuni* were the most contiguous, resolving chromosomes and plasmids that short reads alone missed (Neal-McKinney et al., 2021). By contrast, Illumina short reads (~250 bp) provide very low per-base error rates (mean Q-scores often 30, i.e. 0.1% error). In practice, Illumina-only assemblies of bacteria are highly accurate but fragmented, whereas long-read-only assemblies are complete but error-prone. A long-read-first hybrid approach (long-read assembly followed by polishing with short reads) produces sequences that are both complete and highly accurate (Wick & Holt, 2022). In one analysis of outbreak *Salmonella* isolates, only pipelines that integrated both ONT-polishing and Illumina-polishing obtained near-perfect (>99.99%) accuracy (Luan et al., 2024). Therefore, the most robust method for obtaining completed genomes at the moment is hybrid assembly.

1.4.3 Applications in Long Read Sequencing Enteric Pathogen Surveillance

1.4.3.1 Plasmid detection and chromosomal integration

A 2021 long-read study of 134 multidrug-resistant *Salmonella enterica* isolates covering 33 serotypes used PacBio sequencing to close 233 plasmids, identify large genomic islands (such as SGI-1), and uncover chromosomal insertions of IncQ resistance plasmids in serotype I 4,[5],12:i:– strains (C. Li et al., 2021). Long reads can show mobile resistance elements and chromosomal integration events crucial for AMR surveillance, which short reads cannot. In another case utilising ONT sequencing, researchers found that *Salmonella* Typhi in India separately acquired cephalosporin-resistance genes on various plasmid backbones (e.g., IncX3, IncN), distinct from the IncY plasmid in Pakistan's XDR outbreak (Jacob et al., 2021). Long reads here were vital for resolving plasmid structure and gene context, informing risk assessments of emerging resistance.

1.4.3.2 Complete genome assembly and plasmid closure in foodborne pathogens

A 2019 study used ONT sequencing to obtain complete assemblies of *Salmonella Bareilly* and *E. coli* O157:H7 genomes including their plasmids with >99.9% accuracy within a 4-hour sequencing run (Taylor et al., 2019). This is significant because it demonstrates that ONT sequencing can generate high-quality closed genomes in near real time. This enables the identification of serotypes, virulence genes, and AMR markers.

1.4.3.3 Structural variation and AMR gene copy number in *Campylobacter jejuni*

A comparative investigation of field *C. jejuni* isolates using hybrid Illumina and MinION assemblies revealed that hybrid data enhanced assembly contiguity, enabling chromosome closure, and detected a plasmid in one sample (Neal-McKinney et al., 2021). Large genomic rearrangements, repeating rRNA and tRNA operons, and gene variations that were missed by short-read techniques were all discovered using long-read data. Several *Campylobacter* isolates were found to have complicated variations and extra copies of the *tet(O)* tetracycline resistance gene, which is found both chromosomally and on plasmids, according to 2024 surveillance research conducted in Germany using hybrid genome assemblies. Short-read plasmid prediction algorithms partly failed to identify *tet(O)* and *aadE*, when the genes were present as duplicate or homologous gene variants (Zarske et al., 2024). This emphasises how long reads can accurately place AMR genes and resolve gene copy number, key for understanding resistance potential.

1.4.3.4 Structural variation in *Salmonella*

Long-read sequencing was used in a 2022 study to examine *Salmonella*'s genomic dynamics, revealing new information about AMR and mobile genetic elements. The study identified a novel phage-plasmid hybrid structure carrying multiple resistance determinants, underscoring the capability of long reads to resolve complex genomic architectures that are often missed by short-read approaches (Greig et al., 2022). In addition to expanding our understanding of *Salmonella*'s AMR properties, this

technology allows for unique discoveries such as the finding of P1-bacteriophage-like plasmids, while also giving critical information for monitoring the spread of AMR and guiding public health efforts.

1.4.4 Future Outlook

As sequencing costs continue to decline and base calling accuracy improves, long-read technologies are poised to become a central tool in routine public health microbiology. Their ability to generate comprehensive, high-resolution genomic data with low bias holds great potential for emerging pathogen surveillance, particularly in resource-constrained situations where quick diagnosis and genomic epidemiology are critical.

1.5 DNA Extraction and Sequencing Workflow Optimisation for Pathogen Analysis

WGS has become a crucial method in the investigation of bacterial pathogens, notably enteric bacteria that cause gastrointestinal illnesses. WGS provides high-resolution genomic data, allowing for precise strain identification, detection of virulence and antibiotic resistance genes, and high-resolution epidemiological tracking of outbreaks. To benefit from these advantages, laboratory processes from DNA extraction to sequencing must be optimised for speed, cost-effectiveness, data quality, and repeatability. A critical step in this process is obtaining high-quality pathogen DNA from various sample types and preparing it for sequencing on different platforms (short-read and long-read sequencers). This outlines the key challenges and best practices in DNA extraction from diverse samples (e.g. pure isolates vs. stool), compares methods to maximise DNA yield and quality for Illumina (short-read) and Oxford Nanopore/PacBio (long-read) sequencing, and highlights strategies to improve data quality and reproducibility in pathogen genomic workflows.

1.5.1 Challenges in DNA Extraction from Different Sample Types

Pathogen DNA may be obtained from relatively clean cultured isolates or directly from complex clinical samples like stool. Pure isolates provide a simpler template, but even here, cell wall differences pose obstacles; for example, Gram-positive bacteria have thick peptidoglycan cell walls that are difficult to lyse, often requiring severe

mechanical or enzymatic breakdown (Fernández-Pato et al., 2024). In contrast, stool samples are inherently more challenging despite their high microbial load (Kazantseva et al., 2021). Stool contains a heterogeneous mixture of bacteria (with varying cell wall resilience), human host cells, undigested food, and a variety of PCR-inhibiting substances. Common inhibitors in faecal matter include complex polysaccharides, bile salts, lipids, and urate, all of which can interfere with enzymatic reactions (Srirungruang et al., 2022). These inhibitors can suppress PCR amplification entirely leading to false negatives if not removed. Inhibitors can also affect library preparation efficiency leading to poor ONT sequencing. Thus, an extraction protocol that works well for cultured isolates may falter when applied to stool without additional inhibitor removal steps.

Effective lysis is critical for accurate metagenomic profiling of stool, which contains a taxonomically complex microbiota and many PCR inhibitors. Mechanical disruption (e.g. bead-beating) is highly efficient at lysing tough Gram-positive cells and is widely used to maximise DNA yield and species richness (Isokääntä et al., 2024). In contrast, chemical lysis methods alone often under-represent Gram-positives, while enzymatic treatments (e.g. lysozyme, proteinase K) offer targeted digestion but may not fully disrupt all taxa (Yang et al., 2020). Comparative studies have consistently shown that including bead-beating enhances the detection of Gram-positive organisms and increases diversity metrics (Kwa et al., 2024; Purushothaman et al., 2024; Yang et al., 2020). However, aggressive lysis can shear DNA and release inhibitors, potentially skewing downstream analyses if not paired with adequate purification (Gand et al., 2023). Therefore, the choice of lysis method directly shapes the apparent microbial community and must be tailored to balance coverage and DNA integrity.

The lysis method influences not only which taxa are recovered but also the quality of DNA for sequencing. Short-read sequencing platforms like Illumina are relatively tolerant of fragmented DNA, so protocols that prioritise comprehensive lysis even at the cost of some shearing are acceptable (Becker et al., 2016). In contrast, long-read platforms such as Oxford Nanopore and PacBio require HMW DNA; here, enzymatic or gentle lysis methods are favoured to preserve fragment length (Maghini et al., 2021). Mechanical lysis, while efficient, can compromise long-read performance by producing excessively short DNA fragments, whereas enzymatic protocols yield longer

DNA but may require more time and optimisation to achieve broad coverage. For optimal metagenomic sequencing outcomes, particularly with long-read technologies, protocols must be carefully selected or adapted to preserve both yield and DNA length.

Each lysis procedure includes biases; mechanical lysis risks DNA shearing, chemical methods under-represent robust organisms, and enzymatic protocols may not fully lyse different species if employed alone. These biases can affect pathogen detection, alpha diversity (diversity within a single sample) estimates, and comparative microbiome studies (Kazantseva et al., 2021). Reproducibility also hinges on consistent lysis performance: protocol variations can lead to significant shifts in community composition, especially in the representation of Gram-positive species. To minimise technical variation, standardised protocols combining mechanical and enzymatic lysis, along with effective inhibitor removal, are increasingly recommended (Fernández-Pato et al., 2024; Fiedorová et al., 2019; Maghini et al., 2021; Purushothaman et al., 2024). Harmonising extraction workflows across samples and studies improves the comparability and reliability of metagenomic data, making lysis optimisation a cornerstone of stool-based pathogen genomics and microbiome research(Fiedorová et al., 2019; Kazantseva et al., 2021).

Another challenge in direct clinical samples is the high background of host DNA or other contaminants. For enteric infections in stool, human DNA from shed intestinal cells can substantially exceed bacterial DNA, diluting the pathogen signal while wasting sequencing capacity on host reads. Therefore, before sequencing, several procedures incorporate steps to deplete host DNA (such as selectively lysing host cells or enzymatically digesting human DNA); nevertheless, these steps must be weighed against the additional complexity and expense (T. Charalampous et al., 2019). For example, host DNA is not always present at a significant quantity in stool samples from infected individuals, therefore some projects may view host depletion as an unnecessary expense. Reagent and environmental contamination can be an issue for low-biomass samples. Extraction kits have been shown to introduce contaminating DNA, which can obfuscate results if not managed (Fiedorová et al., 2019). Best practices include processing negative extraction controls (blank samples) alongside real samples to monitor for contamination (Wegl et al., 2021). In summary, the

extraction stage must overcome a number of obstacles, such as rupturing resistant cells and eliminating impurities and inhibitors, all the while preserving DNA integrity, which is particularly important for long-read applications.

1.5.2 High-Quality DNA Yield for Short-Read vs Long-Read Sequencing

Different sequencing platforms have distinct input DNA requirements, so optimising DNA yield and fragment length is key to taking full advantage of each technology. Illumina short-read sequencers (e.g. HiSeq/NovaSeq, NextSeq, MiSeq) typically produce reads of 150–300 bp in length, and their library preparation protocols involve DNA fragmentation (mechanical shearing or tagmentation) to generate these short inserts. As a result, extremely high molecular weight DNA is not required for Illumina sequencing and DNA can be sheared during library prep to the desired size. In fact, Illumina workflows can tolerate somewhat fragmented DNA, and they often include PCR amplification steps that allow successful library construction from relatively low input amounts (tens of nanograms of DNA) (Ribarska et al., 2022). Nevertheless, DNA purity and absence of inhibitors remain critical. Short-read libraries prepared from impure DNA may suffer from amplification biases or even failure of adapter ligation/PCR. For Illumina, a “high-quality” DNA prep means one with moderate fragment size (~ >5–10 kb fragments are usually sufficient) and high purity (A260/A280 ~1.8) (Becker et al., 2016). When working with very limited DNA or with extraction methods that yield small fragment sizes, Illumina kits with built-in PCR can rescue the library, but at the expense of potentially skewing representation of genomic regions (Lou et al., 2021). Therefore, even for short-read sequencing, maximising yield and purity improves consistency. Ideally, one should aim for a DNA input that comfortably exceeds the minimum (to avoid extra amplification cycles) and has optical density ratios of A260/A230 > 2.0, indicating clean DNA with minimal organic contaminants.

In contrast, long-read sequencing technologies (Third-Generation Sequencing) such as Oxford Nanopore Technologies (ONT) and Pacific Biosciences (PacBio) have performance that directly hinges on DNA length and integrity. These platforms can sequence DNA fragments tens of kilobases long, and their power lies in reading long contiguous segments, but only if the input DNA is not already degraded. For ONT (e.g. MinION, GridION, PromethION), typical library preparation by ligation recommends ~1 µg of high molecular weight genomic DNA (with fragment lengths ideally 50 kb or more)

for optimal results (Maghini et al., 2021). Starting with very large DNA enables obtaining ultra-long reads (in some cases >100 kb) which are valuable for resolving repetitive regions and complete genomes. If DNA is heavily sheared (e.g. mostly <10 kb fragments), nanopore sequencers will still produce reads, but they will be shorter and a lot of sequencing yield may be "wasted" on very short fragments that don't contribute much new information. PacBio HiFi sequencing (circular consensus sequencing) similarly benefits from HMW DNA . Although PacBio's HiFi libraries often target an insert size of ~15–20 kb, the recommendation is to start with genomic DNA averaging >40 kb in length (Bronner et al., 2025). Having ultra-HMW input allows the DNA to be sheared to the desired 15–20 kb size while ensuring the DNA wasn't already badly fragmented. Studies from PacBio note that starting with HMW DNA improves read length and yield, whereas degraded DNA can result in shorter reads and lower throughput (Pacific Biosciences, 2022). Overall, long-read platforms demand more from the extraction process, DNA must be not only pure but also as intact as possible.

To achieve high yields of intact DNA, researchers often turn to gentler extraction techniques or specialised kits. Traditional phenol–chloroform extraction is known for yielding high-purity DNA with large fragment sizes and minimal reagent cost (Wright et al., 2017). Indeed, phenol-chloroform is often considered a gold-standard for DNA purity and length (avoiding the silica membrane or beads that might shear DNA), though it is labour intensive and uses toxic reagents (Chachaty & Saulnier, 2000). Many laboratories prefer column-based kits for convenience and safety. However, not all kits are equal when it comes to HMW DNA: for example, recent novel technologies like Nanobind disks (a silica-coated magnetic disk method) tend to recover significantly larger DNA fragments than conventional silica spin columns or magnetic bead (Liu et al., 2019). The latter methods involve passing DNA through membranes or frequent pipetting, which can introduce mechanical shear and break long DNA (Quick & Loman, 2019). By minimising such forces, HMW extraction kits can routinely produce genomic DNA tens to hundreds of kilobases in length. This is especially important for long-read sequencing of bacterial pathogens when one wants to assemble complete genomes or plasmids short DNA would negate the advantage of long-read sequencers. Another best practice is to include proteinase K digestion during lysis and avoid harsh conditions that might damage DNA (Gautam, 2022). Proteinase K (or similar proteases) helps inactivate nucleases present in the sample that could otherwise chew up DNA.

In addition, avoiding excessive vortexing or repeated freeze-thaw cycles preserves DNA length (Trigodet et al., 2022). When working with tough samples a combination of methods may be used, a mild mechanical disruption to open cells followed by immediate gentle handling of the lysate to spool or bind intact DNA (Barbosa et al., 2016; Nadkarni et al., 2009).

It should be noted that maximising DNA length can sometimes conflict with maximising yield. For instance, vigorous bead-beating will crack open all cells but will also shear DNA into smaller fragments. Therefore, when long-read sequencing is the goal, scientists often seek a balance, using enough mechanical or enzymatic lysis to liberate DNA from all organisms, but not so harsh as to fragment all the DNA. Enzymatic lysis (e.g. lysozyme for Gram-positives, alongside gentle SDS/proteinase K) followed by careful extraction can sometimes replace extreme bead-beating when ultra-long DNA is needed (Waters et al., 2022). If mechanical disruption is unavoidable for certain tough bacteria, researchers might size-select the output (for example, using a pulsed-field gel or a size-selection magnetic bead protocol) to remove the bulk of small DNA fragments before library prep (Huptas et al., 2016)/. In summary, short-read sequencing workflows are relatively forgiving with DNA fragment size and input amount (as long as inhibitors are removed), whereas long-read workflows demand more optimisation of the extraction method to produce high-molecular-weight, inhibitor-free DNA for successful sequencing.

1.5.3 Best Practices for Data Quality and Reproducibility

Ensuring data quality and reproducibility in pathogen genomics starts with standardising the sample storage and preparation. Variation introduced at the DNA extraction stage can lead to significant downstream biases, as noted above. Therefore, one key best practice is the harmonisation of protocols across all samples and, if possible, across laboratories in a study. Using the same extraction kit and method for all samples (with consistent input amounts, incubation times, etc.) reduces technical variability. In practical terms, labs often validate a few different extraction methods and then adopt the one that gives the best yield/quality for their sample type as a standard operating procedure. Initiatives like the International Human Microbiome Consortium have even recommended standardised protocols (e.g. International Human Microbiome Standards protocol Q for stool DNA extraction) that were shown to

perform well across multiple criteria (Dore et al., 2015). Adhering to such standardised protocols can improve reproducibility and allow comparisons of data between studies with greater confidence. An ongoing challenge in this field is maintaining pace in an environment of continuous innovation, where research is constantly analysing and proposing improved methods (Rintarhat et al., 2024).

After extraction, rigorous quality control (QC) is essential. It is good practice to quantify DNA using a sensitive, specific method like a Qubit fluorometer (which measures double-stranded DNA concentration) and to evaluate purity by spectrophotometry (NanoDrop). Pure DNA typically shows an A_{260}/A_{280} ratio around 1.8, indicative of low protein/phenol contamination, and A_{260}/A_{230} above ~2.0, indicative of low humic acid, carbohydrate, or salt carryover (Reuter & Zaheer, 2016). Because impure DNA can result in lower sequencing throughput or quality failures, it is best to repurify a sample that is not extremely pure before sequencing. Additionally, checking DNA integrity by running a portion on an agarose gel or using capillary-based electrophoresis (e.g. TapeStation) can confirm the fragment size distribution. Consistently performing these QC stages on each batch guarantees that only acceptable DNA enters the library preparation, improving the consistency of sequencing results.

Implementing control samples in the workflow bolsters confidence in the data. Negative controls (extraction blanks) help detect any background DNA contamination introduced during the process (Fiedorová et al., 2019; Salter et al., 2014). Positive controls (e.g. a known quantity of a reference organism spiked into a subset of samples, or a reference DNA sample included in each batch) can serve as an internal check on DNA recovery and sequencing performance. If the known control's results fluctuate or drop, that signals an issue with that batch's extraction or sequencing. Moreover, performing replicate extractions on the same sample (when material is plentiful) is a way to gauge method consistency highly reproducible workflows should yield similar results from replicates. This was highlighted by observations that technical variation from different DNA extraction methods can sometimes be as large as or larger than biological variation (Kazantseva et al., 2021). By minimising such technical variation through protocol consistency and controls, the data will more reliably reflect true biological differences rather than artifacts.

Another best practice for reproducibility is automation of DNA extraction and library prep where feasible. Automated extraction systems (using robotic liquid handlers or dedicated instruments like QIAcube or KingFisher) can reduce person-to-person variability and handling errors. According to studies, automated techniques can increase consistency and throughput while producing DNA yields and quality that are on par with manual techniques (Fernández-Pato et al., 2024). Automation, when combined with adequate calibration and maintenance, helps to ensure that each sample is processed under the same conditions, boosting repeatability in large projects or clinical labs with a high volume of samples.

1.5.4 Speed and Cost Considerations in Workflow Optimisation

When optimising workflows, speed and cost-efficiency are often key drivers, especially in clinical or public health settings where time-to-result and budget are critical. There is usually a trade-off between the fastest possible method and the one that yields the absolute highest quality data. In traditional pathogen analysis, a stool sample is first cultured to isolate the pathogen, and then DNA is extracted from the isolate for sequencing. This culture-based approach has the advantage of enriching the target organism and is cost-effective in terms of sequencing as only the pathogen genome is sequenced. However, culture steps are time-consuming often requiring over 24 hours just to grow the colonies, and some pathogens may not grow well in the lab at all. Furthermore, it's now recognised that as culture can be less sensitive than molecular detection, a significant fraction of infections might be missed by culture due to overgrowth by other flora or stringent growth requirements, leaving up to ~80% of cases unresolved in some studies (Peterson et al., 2022). In recent years, culture-independent metagenomic sequencing of stool has emerged as a modern alternative. By directly extracting DNA from stool samples and performing sequencing whether targeting specific genes or conducting whole metagenome shotgun sequencing this method can significantly shorten the time to diagnosis. It enables rapid DNA extraction and sequencing, potentially within the same day bypassing the need for cultivation. Moreover, it detects a wide spectrum of pathogens, including those that may escape detection by conventional culture methods or diagnostic panels (Peterson et al., 2022). Metagenomic sequencing has potential to simultaneously provide diagnostic identification, antimicrobial resistance genes, and subtyping data from a sample.

Information that traditionally would require separate culture and typing steps. The trade-offs of direct sequencing include: (1) higher sequencing cost per sample because one must sequence through host and microbiome DNA, and (2) the need for powerful data analysis pipelines to fish out pathogen sequences from background. Encouragingly, the cost of shotgun metagenomics has been dropping and is reported to be approaching the combined cost of traditional testing which might involve a multiplex PCR panel, reflex culture, and then WGS of the isolate(Peterson et al., 2022). As sequencing cost and speed continue to improve, it is conceivable that a single sequencing-based test could replace the multi-step workflows in the near future.

The choice between Illumina, ONT, and PacBio also has implications for speed and cost. Illumina platforms, especially high-throughput models, remain the most cost-effective for large projects. The cost per gigabase of sequence data has fallen below £50 on instruments like NovaSeq. This makes Illumina very attractive for sequencing many bacterial genomes or doing deep sequencing for metagenomics, as the per-sample cost can be low when multiplexing many samples. The trade-off is that Illumina runs are not as rapid in turnaround: a typical run can take 1–2 days (plus library prep time), and results are only available after the run completes. Oxford Nanopore sequencing offers a different model: relatively low capital cost for the device and the ability to sequence in real-time. An ONT MinION flow cell can be used for a single sample if needed, and data streams off the device immediately as DNA is sequenced. This has enabled scenarios like near-real-time genomic surveillance, where initial results (e.g. detection of a pathogen or key resistance gene) can be obtained within hours of starting sequencing. From a cost perspective, ONT's consumable cost per sample can be higher for small projects (one flow cell per sample might cost a few hundred pounds) but scales favourably for larger flow cells (e.g. PromethION can sequence many samples on one flow cell). Importantly, nanopore sequencing's portability and speed (no need to wait for a batch or a scheduled run) make it ideal for rapid field deployments or urgent clinical analyses. PacBio sequencing, particularly with the Sequel II/IIe or new Revio system, has carved a niche for projects requiring highly accurate long reads. PacBio runs are generally slower (a HiFi run might take ~15-30 hours) and the instruments are expensive to operate, but the data quality (HiFi accuracy >99.9% on long reads) is exceptional for applications like complete genome assembly. PacBio's cost per base has historically been higher

than Illumina's, though the introduction of the Revio (with much higher throughput per run) is bringing those costs down. In practice, laboratories may use a hybrid approach: Illumina for routine high-throughput screening of many isolates (cheap and accurate for single nucleotide variants), and ONT or PacBio for select samples where long-read data is needed (e.g. to resolve plasmid structures or repeat elements). Each lab must optimise based on their specific needs, if speed is paramount (for example, in a hospital outbreak scenario), ONT might be favoured. If cost per sample is the limiting factor and hundreds of genomes need sequencing, Illumina is often the choice.

Beyond the sequencing platform, cost efficiency can be improved by miniaturising protocols, batch processing and automation. Preparing samples in batches saves setup time and makes better use of consumables. Miniaturising protocols increasing the number of samples that can be produced with a given set of consumables. As mentioned, automated extraction or library prep can reduce labour costs and free up personnel. Commercial kits are more expensive per sample but they save time and typically produce cleaner DNA with consistent yields. Thus, many labs choose kits for their convenience and reliability, despite the higher per-sample cost, especially when labour costs and the value of reliable results are considered.

1.6 Aims, Objectives, & Research Questions

Aims:

- To develop and implement a robust, semi-automated protocol for HMW DNA extraction from bacterial isolates and human stool using the Fire Monkey kit on the Tecan A200 platform, optimised for long-read sequencing applications.
- To provide a knowledge base for *Salmonella* diversity within individual patients suffering from gastroenteritis.
- To evaluate the impact of different stool storage conditions on the recovery and genome quality of *Campylobacter* using metagenomic sequencing.
- To assess the utility of long-read metagenomic sequencing for culture-free detection and strain-level characterisation of *Campylobacter* directly from human stool.

Objectives:

- Adapt and validate protocols Fire Monkey for *Salmonella* isolates and complex stool samples using the Tecan A200 semi-automated robotic system.
- Benchmark DNA yield, purity, and fragment length against a commercial extraction system to achieve DNA suitable for long-read sequencing.
- Isolate 20 *Salmonella* colonies per patient, perform whole-genome sequencing, and apply bioinformatic analyses to assess the genetic diversity, sequence types, and presence of antimicrobial resistance and virulence genes among recovered isolates.
- Compare the effectiveness of three storage methods over time, using culture, qPCR, and metagenomic approaches to assess *Campylobacter* viability, DNA integrity, and sequencing outcomes.
- Compare the performance of two stool DNA extraction methods (Fire Monkey and Promega Maxwell) for Oxford Nanopore long-read sequencing, evaluating their ability to reconstruct high-quality *Campylobacter* genomes based on assembly quality, sequence typing, antimicrobial resistance detection, and concordance with isolate-derived genomes.

Research Questions:

Does sequencing multiple colonies per patient reveal greater within-host genomic diversity of *Salmonella* compared with conventional single-colony sequencing?

Hypothesis: Sequencing multiple colonies from the same infection may reveal greater genetic variation than single-colony approaches.

To what extent does stool preservation method (raw freezing, glycerol freezing, DNA/RNA Shield) affect the detectability and genomic completeness of *Campylobacter* recovered by metagenomic sequencing over time?

Hypothesis: Different preservation approaches may variably maintain nucleic acid integrity, with implications for how faithfully genomic data reflect the original sample composition.

How does the quality and integrity of high-molecular-weight (HMW) DNA extracted from stool samples influence the success of long-read metagenomic recovery of *Campylobacter* genomes?

Hypothesis: Improved DNA quality is expected to enhance the recovery and resolution of pathogen genomes from complex stool samples.

This project was carried out at Quadram under an iCASE studentship made by the Medical Research Council (MRC) through the Doctoral Antimicrobial Research Training (DART) MRC iCase Programme. This project linked RevoluGen, a leader in the field of HMW DNA extraction techniques with Quadram's enterprise in long read sequence analysis. As part of the project, the Tecan A200 robotic system was loaned to Quadram, while 96-well filter column plates and Fire Monkey reagents were provided as consumables. I would like to express my gratitude to Dr. Georgios Patsos for his support during the development of the Fire Monkey processes. I would also like to thank Dr. Rebecca Entwistle and Dr. Helena Patsos for coordinating the shipment of reagents throughout the project.

2 Automating High Molecular Weight DNA Extraction: Fire Monkey Protocols for bacterial isolates and stool on the Tecan A200

Chapter contributions: Dr Georgios Patsos developed the foundational Fire Monkey protocol for the Tecan A200.

Methods developed in this chapter have been utilised in publications:
Rudder, S. J., Djeghout, B., Elumogo, N., Janecko, N., & Langridge, G. C. (2025). Genomic diversity of non-typhoidal *Salmonella* found in patients suffering from gastroenteritis in Norfolk, UK. Microbial Genomics, 11(8), 001468. https://doi.org/10.1099/mgen.0.001468

Carter, C., Hutchison, A., *Rudder, S*., Trotter, E., Waters, E. V., Elumogo, N., & Langridge, G. C. (2023). Uropathogenic Escherichia coli population structure and antimicrobial susceptibility in Norfolk, UK. *Journal of Antimicrobial Chemotherapy*, *78*(8), 2028-2036. https://doi.org/10.1093/jac/dkad201

2.1 Introduction

HMW genomic DNA is a critical starting material for long-read sequencing technologies (Jaudou et al., 2022; Trigodet et al., 2022). Platforms like ONT can create reads as long as the input DNA fragments. The current record exceeding 4 million bases in one read(Eagle et al., 2023). By using intact HMW DNA, reads that span repetitive or challenging sections can be sequenced, making it easier to assemble entire genomes and identify structural variations. In bacteria, the longest repeating areas are frequently the ~5-7 kb rRNA operons. Reads that cross these regions and anchor in the surrounding DNA are highly desirable. 20 kb sequencing reads are advised as a target to ensure complete genome assembly (Cao et al., 2017; Koren & Phillippy, 2015; Wick et al., 2023). In addition to fragment length, yield is a crucial consideration. ONT's library preparation kits generally require hundreds of nanograms to micrograms of input DNA, on the order of 400-1000 ng (Eagle et al., 2023). Obtaining pure HMW DNA maximises read length and assembly quality in ONT sequencing (Gand et al., 2023; Kruasuwan et al., 2024). Consequently, robust protocols for extracting

large, intact DNA at high yields have become essential, particularly as long-read sequencing is applied to both isolated bacterial genomes and complex metagenomic samples. This chapter explores the evolution of a semi-automated HMW DNA extraction method on a Tecan A200 platform utilising Fire Monkey protocols. It demonstrates successful methods for extracting high-yield, high-integrity DNA from bacterial isolates and stool samples.

Numerous HMW DNA kits and techniques are available for sequencing bacterial isolates, and each one uses a different methodology. Chemical or enzymatic lysis combined with DNA capture using magnetic beads or silica spin columns is the most common technique for cultured bacterial cells. Typically, these procedures yield between 5 to 15 µg of DNA, with DNA fragments reaching sizes up to 300 kb. Notably, more than 60% of these DNA fragments are ≥20 kb in length (see Table 2.1). It's important to keep in mind that the particular bacterial strain influences the extraction method selection because some bacteria present more difficulties with regard to lysis efficiency (Danaeifar, 2022; de Bruin et al., 2019).

Various methods and protocols are available for DNA extraction from stool samples, with bead beating followed by DNA capture in silica filter columns being the most prevalent approach. Alternative procedures, such as chemical or enzymatic lysis, can be used to create HMW DNA without the use of beads. Silica filters, magnetic beads, or genomic tips are commonly used to extract DNA from stool samples. While DNA yields can exceed 15 µg, it is common to obtain 1-2 µg of DNA per 0.1-0.5 g of input material (see Table 2.2). Reported HMW DNA fragment sizes from human stool samples typically range from 4 to 50 kb (LeFrançois & Cunningham, 2019; Maghini et al., 2021; Purushothaman et al., 2024).

The Promega Maxwell RSC, the Bioer GenePure Pro, and the KingFisher Apex System are three semi-automated devices that have been tested and validated to extract DNA from human stool samples. These devices effectively capture and purify nucleic acids employing magnet beads for DNA recovery and have been acknowledged for their capacity to recover DNA at yields comparable to manual kits, while greatly lowering hands-on time and enhancing workflow efficiency (Kwa et al., 2024). In addition to these systems, the Fire Monkey kit, employing the Tecan A200 positive pressure

system, offers an alternative semi-automated approach using a 96-position filter column plate. In this chapter I test and implement the 96-well format Fire Monkey kit using the Tecan A200 and develop a Fire Monkey human stool HMW DNA preparation.

Table 2.1: Approximations of DNA Yield and Fragment Size of Commercial HMW DNA Extraction Kits Targeting Bacterial Cells

Method	Principle	DNA Yield	Fragment Size	Refs
Qiagen DNeasy	Silica spin column	~20 ng/µL (2 µg total)	~87% ≥20 kb fragments	(Eagle et al., 2023)
Qiagen EZ1 DNA Tissue	Magnetic silica beads	~42 ng/µL (4 µg total)	~91% ≥20 kb fragments	(Eagle et al., 2023)
Lucigen MasterPure	Precipitation	~62 ng/µL (6 µg total)	~62% ≥20 kb fragments	(Eagle et al., 2023)
MasterPure (In-house mod)	Enzymatic + precip.	~59 ng/µL (6 µg total)	~97% ≥20 kb fragments	(Eagle et al., 2023)
Omega E.Z.N.A. Bacterial	Silica spin column	~78 ng/µL (7–8 µg total)	~66% ≥20 kb fragments	(Eagle et al., 2023)
Qiagen Genomic-tip 20/G	Anion-exchange gravity	~10 µg	20-250 kb	(Becker et al., 2016)
Qiagen MagAttract HMW	Magnetic beads + SDS lysis	~10 µg	15-300 kb	(Becker et al., 2016)
Zymo Quick-DNA HMW MagBead	Magnetic beads + enzymes	5–15 µg	~50 kb	www.zymoresearch.eu
NEB Monarch HMW	Glass beads + gentle lysis	5–15 µg	50–250 kb	www.neb.com
Circulomics Nanobind	Nanobind disk (silica)	5–25 µg	50–300 kb	www.pacb.com
Promega Maxwell RSC Cultured Cells DNA	Magnetic beads (cartridge-based)	5–15 µg	~20–100 kb	www.promega.co.uk
RevoluGen Fire Monkey	Silica spin column	5–15 µg	100–130 kb	www. revolugen.co.uk

(Note: Yields and fragment sizes can vary with input amount and handling; values above are from referenced studies or manufacturer specs and input varies across studies)

Table 2.2: Approximations of DNA Yield and Fragment Size of Commercial HMW DNA Extraction Kits Targeting Human Stool

Method	Principle	DNA Yield	Fragment Size	Refs
Zymo Quick HMW MagBead kit	Enzymatic + MagBeads	~1.5 µg (0.25 g stool)	~50 kb peak	www.zymoresearch.com
QIAamp PowerFecal Pro DNA Kit	Bead-beating + Silica	~10-17 µg (0.5 g stool)	~20 kb	(LeFrançois & Cunningham, 2019)
QIAamp PowerFecal Pro DNA Kit	Bead-beating + Silica	~4.9 µg (1 mL eSwab)	4392 bp (read N50)	(Purushothaman et al., 2024)
QIAamp DNA Mini Kit	Enzymatic + Silica	~4.4 µg (1 mL eSwab)	7152 bp (read N50)	(Purushothaman et al., 2024)
ZymoBIOMICS DNA Miniprep	Bead-beating + Silica	~2-15 µg (0.5 g stool)	peaks ~5 kb	(LeFrançois & Cunningham, 2019)
Maghini et al. 2021 protocol	Enzymatic + Phenol/Tip	1–2 µg (0.3-0.5 g stool)	15–50 kb	(Maghini et al., 2021)
Maxwell RSC Fecal Microbiome DNA Kit	Chemical + MagBeads	1–3 µg (0.1-0.3 g stool)	Peak length tens of kb	www.promega.co.uk
Maxwell® RSC Buccal Swab DNA Kit	Enzymatic + MagBeads	~12.25 µg (1 mL eSwab)	7893 bp (read N50)	(Purushothaman et al., 2024)
Maxwell® RSC Cultured Cells DNA Kit	Enzymatic + MagBeads	~4.5 µg (1 mL eSwab)	6321 bp (read N50)	(Purushothaman et al., 2024)

(Note: Yields and fragment sizes can vary with input amount and handling; values above are from referenced studies or manufacturer specifications and input varies across studies)

2.2 Aims and objectives

The work outlined in this chapter aimed to:

- Implement the Fire Monkey HMW DNA extraction preparation using the Tecan A200 system
- Develop a protocol for HMW DNA extraction from clinical *Salmonella* isolates
- Develop a protocol for HMW DNA extraction from stool

2.3 Materials and methods

2.3.1 DNA quantification – Single tube assay

The Qubit™ dsDNA BR Assay Kit (Q32853, Thermo Fisher, UK) was used as follows: 199 µL of Qubit dsDNA BR buffer and 1 µL of Qubit™ dsDNA BR Reagent were combined to prepare a master mix of the appropriate volume. For standards, 190 µL of the master mix was mixed with 10 µL of the Qubit™ dsDNA BR Standards supplied with the kit. For samples, 198 µL of the master mix was mixed with 2 µL of DNA. Each sample was vortexed for 10 seconds and allowed to rest for at least 2 minutes before being measured using a Qubit™ 3.0 Fluorometer. All standards and samples were quantified using Qubit™ assay tubes (Q32856, Thermo Fisher, UK). During Oxford Nanopore Technologies (ONT) library preparation (1.2.7.1) 1 µL of DNA library was used with 199 µL of master mix.

2.3.2 DNA quantification – Plate assay

The Quant-iT™ dsDNA Assay Kit (Q33130, Thermo Fisher, UK) was used as follows: 199 µL of Quant-iT™ dsDNA BR buffer and 1 µL of Quant-iT™ dsDNA BR reagent were combined to prepare a master mix of the appropriate volume. For standards, 190 µL of the master mix was mixed with 10 µL of the λ dsDNA BR standards (0, 5, 10, 20, 40, 60, 80, and 100 ng/µL) supplied with the kit. For samples, 198 µL of the master mix was mixed with 2 µL of DNA. All standard and samples were added to a CytoOne flat bottom, non-treated 96-well plate (CC7672-7696, Starlab, Germany). The plate was gently vortexed, briefly centrifuged and allowed to rest for at least 2 minutes. Readings were taken using a Promega GloMax Discover System (Promega, USA).

2.3.3 DNA sizing

DNA integrity and size were estimated using the Genomic DNA ScreenTape analysis (5067-5365 & 5067-5366, Agilent Technologies, USA) on an Agilent TapeStation. Each sample was prepared by mixing 1 µL of genomic DNA with 10 µL of Genomic DNA Sample Buffer in a PCR tube. For each assay, 1 µL of Genomic DNA Ladder was mixed with 10 µL of Genomic DNA Sample Buffer. All samples were gently vortexed and briefly centrifuged prior to analysis. A subset of samples was shipped to RevoluGen Ltd for DNA size analysis using an Agilent Femto Pulse.

2.3.4 Nanodrop

A Nanodrop Spectrophotometer ND-100 (Thermo Fisher, USA) was used to analyse the purity of DNA samples. The device was first engaged by testing 1 µL of water. Once active the system was blanked with 1 µL of RevoluGen Elution Buffer and then DNA samples were tested by adding 1 µL to the device.

2.3.5 Proteinase K

Proteinase K, recombinant PCR grade powder (Roche, Germany) was dissolved in a buffer consisting of 30 mM Tris HCl, 30 mM EDTA, pH 8, to make a solution at 20 mg/mL.

2.3.6 Host depletion reagents

Saponin (Tokyo Chemical Industry UK, UK) was made up at 1% in phosphate buffer saline (PBS) on the day of use. HD buffer (5 M NaCl, 0.1 M $MgCl_2$) was made up in 40 mL batches, filter sterilised and stored at room temperature.

2.3.7 Bead clean

AMPure XP beads (Beckman Coulter, USA) were added at 1x or 0.6x volume of DNA sample. The sample was mixed by brief vortex and then rested at room temperature for 5 minutes. The sample was then placed on a magnetic rack and left for ~ 2 minutes for the beads to attach to the magnet and the solution to become clear. The supernatant

was removed and 500 µL of 70% ethanol was pipetted over the beads. The 70% ethanol was removed, and the washing process was repeated. On the second wash the 70% ethanol was removed by 1000 µL pipette, and then any residue was collected from the bottom of the tube with a 10 µL pipette. The sample was removed from the magnet and 50 µL RevoluGen elution buffer (EB) buffer was added to the sample. The sample was flicked until the beads resuspended and then rested at room temperature for 5 minutes before returning to the magnetic rack. Once the beads had attached to the magnet and the solution was clear the supernatant was collected in clean 1.5 mL Eppendorf tube.

2.3.8 Preparing Polyvinylpolypyrrolidone

Polyvinylpolypyrrolidone (PVPP) (Sigma Aldrich, USA) was dissolved in PBS at 2% w/v. This solution was autoclaved.

2.3.9 Stool collection

Stool specimens surplus to requirements were collected from the National Health Service (NHS) Eastern Pathology Alliance (EPA) laboratory, Norwich, Norfolk, United Kingdom (UK) between March 2020 and August 2022. Three samples were collected before the start of this project in 2020, and five samples were collected during the project. All samples were marked *Salmonella* spp. positive at the EPA, as determined by a PCR-based culture independent testing panel (Gastro Panel 2, EntericBio, Serosep, United Kingdom). Aliquots of up to 20 mL were transferred triple contained to the Quadram Institute Biosciences (QIB) where they stored a 4 °C overnight (15 hours). The next morning samples were split and stored as up to 1 mL aliquots raw and as a 50:50 mix with *Brucella* Broth supplemented with 17.5% glycerol. These aliquots were transferred to the University of East Anglia (UEA) Biorepository where they were stored at -80 °C. *Salmonella* positive stool specimens were stored until a serovar was confirmed by the UK Health Security Agency (UKHSA), this was a safety measure put in place to avoid cultivation of a Hazard group 3 *Salmonella* species. *Campylobacter* positive stool specimens were identified using the same PCR panel array and were free to use upon collection.

2.3.10 Fire Monkey using the A200

Numerous Fire Monkey Tecan A200 protocols are reported in this chapter. The Tecan A200 enables automation of the wash and elution steps. The cell lysis and cleaning of the lysate occurs manually and are described in protocols presented for *E. coli*, *Salmonella,* and stool. The installation of the Tecan A200 at Quadram involved the development of the foundational Fire Monkey HMW DNA extraction protocol by Dr. Georgios Patsos. This protocol includes the washing and elution stages translated from the spin column version of the Fire Monkey HMW DNA extraction kit. The Tecan A200 protocol consists of several key operations: flash, wash, elute, wait and message. Flash operations utilise air pressure to force lysate or wash through the filter columns. Wash and elute operations involve the addition of wash solution or elution buffer to the columns. During wait operations, the robot remains idle, and message operations display messages on the console (Table 2.3).

Table 2.3: Template Protocol for Fire Monkey HMW DNA Extraction Kit Washes and Elution

Step	Operation	Parameter	Volume (µL)	Time (Min)
1	Flash	load ec	-	-
2	Wash	500 µL	500	-
3	Flash	w1 ec	-	-
4	Wash	500 uL	500	-
5	Flash	w2 ec	-	-
6	Flash	QIAamp 96 Viral RNA - Drying 30mins	-	-
7	Message	Place Collection Plate	-	-
8	Elute	100 µL	100	-
9	Wait	10 Min	-	10
10	Flash	elution ec	-	-
11	Message	Place collection plate	-	-
12	Elute	100 µL	100	-
13	Wait	10 Min	-	10
14	Flash	elution 2 ec	-	-
15	Message	Method complete	-	-

The flash operations are points in the protocol where modifications can be made to ensure the lysate or wash pass through the column in allocated time. The two variables that can be modified are pressure and time. Pressure can be increased to help force solution through the column. Time can be increased to give the pressure

applied more seconds to force the solution through the filter. Increases to pressure need to be made gradually to avoid damage to the instrument and column filters. The flash profiles for the foundational protocol can be seen in Figure 2.1. Images of the Tecan A200 set up can be seen in Figure 2.2.

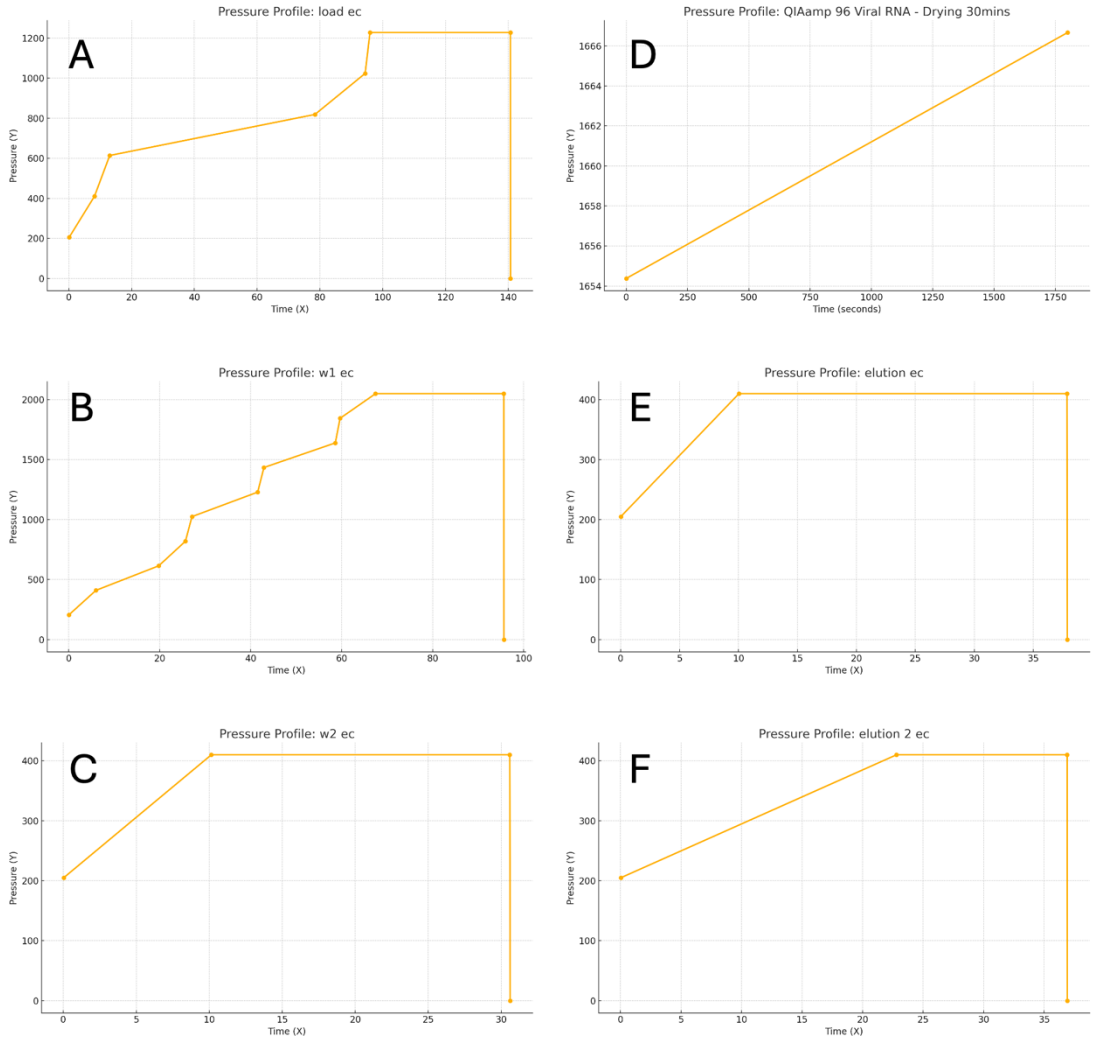

Figure 2.1: Pressure profiles are shown for each programmed *flash* operation used during the Fire Monkey high-molecular-weight DNA extraction workflow performed on the Tecan A200 positive-pressure workstation. Flash operations allow fine control of both applied pressure and dwell time to ensure lysate, wash buffers, and elution buffer pass uniformly through the silica column. These parameters can be adjusted to optimise flow consistency, with pressure increases made conservatively to avoid damaging the instrument or filter units. Panels depict the pressure–time traces for the major protocol stages: A) Step 1, lysate loading; B) Step 3, LSDNA/ethanol wash; C) Step 5, 75% isopropanol wash; D) Step 6, column drying; E) Step 10, first elution; and Step 14, second elution. Time is displayed in seconds. Together, these traces represent the baseline flash profile used as the template for all subsequent optimisation experiments.

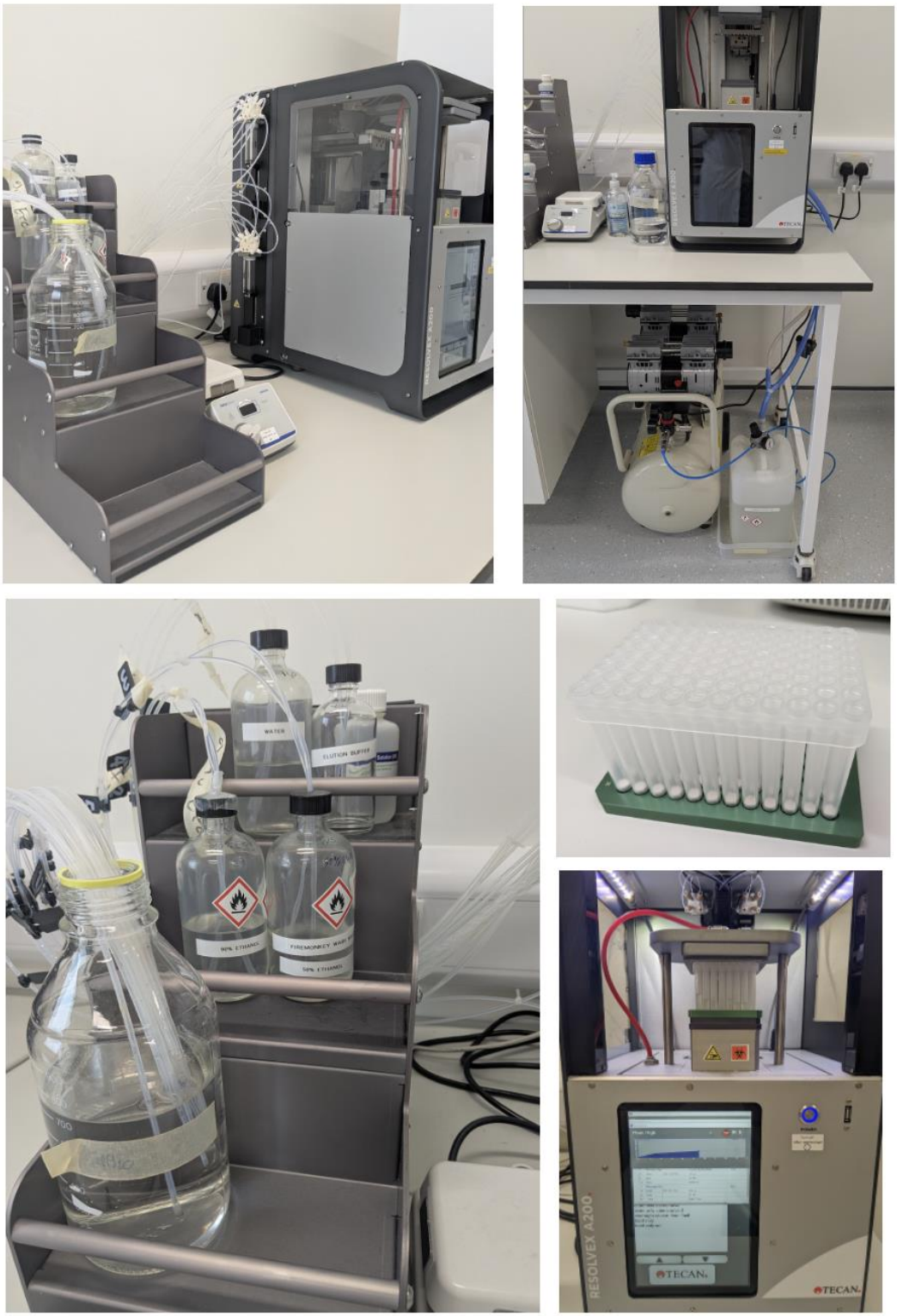

Figure 2.2: Images of the Tecan A200 set up to run the Fire Monkey HMW DNA preparations. Clockwise from top left the images depict side view of A200 with reagent bottle stack with piping into the A200, front view of A200 with compressor and waste container under the desk, 96-well collection plate in bracket, A200 running a flash operation, and the reagent bottle stack.

2.4 Results

2.4.1 Testing the Fire Monkey Tecan A200 platform with *Escherichia coli*

Upon installation at Quadram, the Tecan A200 robot was tested using a set of uropathogenic *Escherichia coli* (UPEC). The UPEC formed part of a collection on isolates studied as part of Cailean Carter's PhD (Carter et al., 2023). The aim of this work was to learn how to use the Tecan A200 and identify a protocol for extraction of HMW DNA from 101 *Escherichia coli* (*E. coli*) isolates.

2.4.1.1 Protocol

Single colonies of UPEC were grown in Bijou containers in 2 mL Luria Broth (LB) for 16 hours in an incubator shaker set to 37°C and 200 rpm. From the overnight cultures, 700 µL of culture was moved to 1.5 mL tubes and pelleted by centrifuge at 16,000 rpm for 1 minute. The supernatant was then removed and discarded. To lyse the cells, 3 mg/µL lysozyme was added to a lysis buffer containing 1.2% Triton X-100, 100 µL of this lysis buffer was added to the pellet. Samples were mixed by 5 pipette mixes and briefly vortexed (10 seconds) before incubating at 37°C for 10 minutes. A master mix of 300 µL LSDNA buffer and 20 µL Proteinase K was prepared for the appropriate number of samples. The 320 µL master mix was added to samples before mixing with 5 pipette mixes and brief vortexing (10 seconds). These samples were then incubated at 37°C for 20 minutes. After incubation, 20 µL RNase A solution was added to the samples, which were then rested at room temperature for 5 minutes. A 350 µL volume of Binding Solution (BS) solution was added to the samples, which were mixed by vortexing (10 seconds). Finally, a 400 µL volume of 75% isopropanol was added to the samples, which were mixed by vortexing (10 seconds). The lysate was loaded into the columns of a Cerex 96-well plate with Cytiva glass fiber filters. The Tecan A200 was run using the methods *E. coli* SR version 2.xml (Table 2.4).

Table 2.4: Tecan A200 operations for *E. coli* SR version 2 protocol

Step	Operation	Parameter	Solvent (v/v)	Volume (µL)
1	Flash	load ec plus 30 s pressure increase		
2	Wash	500 µL	EtOH:WS (50:50)	500
3	Flash	w1 ec		
4	Wash	500 µL	EtOH:H_2O (90:10)	500
5	Flash	w2 ec		
6	Flash	QIAamp 96 Viral RNA - Drying 30 mins		
7	Message	Place Collection Plate		
8	Elute	100 µL	Elution Buffer	100
9	Wait	10 min		
10	Flash	elution ec		
11	Message	Place collection plate		
12	Elute	100 µL	Elution Buffer	100
13	Wait	10 min		
14	Flash	elution 2 ec		
15	Message	Method complete		

Within Step 1 of the A200 protocol the time and pressure were increased to allow lysate from all samples from a set of 48 *E. coli* to pass through the column filter avoiding clogging of filters (Figure 2.3). Clogging can lead to a column overflowing as washes are added or incomplete sample capture after improper washing. Both scenarios are undesirable, especially the overflowing of columns; in a scenario where a clogged column looks set to overflow, I recommend stopping the protocol and removing excess lysate/buffer from the column with a pipette before resuming the protocol.

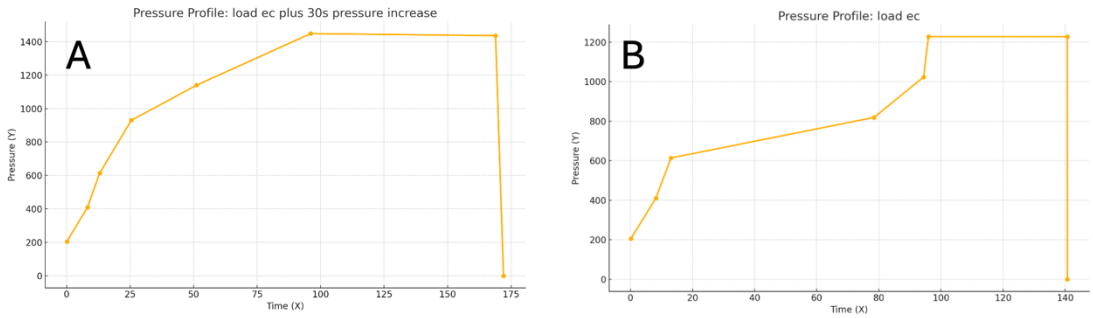

Figure 2.3: Pressure and time settings for Step 1 (lysate loading phase). A, *E. coli* SR version 2 protocol for the Tecan A200. B, Foundational protocol. During Step 1 of the automated protocol, the applied pressure and dwell time were increased to ensure complete passage of lysate through the silica filter for all samples in a 48-sample E. coli extraction set. These adjustments were implemented to prevent column clogging, which can result in incomplete lysate capture or, in severe cases, column overflow during downstream wash steps. Panel A shows the pressure–time profile used in the *E. coli* SR v2 protocol, while Panel B displays the corresponding Step 1 settings in the foundational template protocol.

2.4.1.2 Implementation

DNA was extracted from 101 UPEC with the protocol described above (2.4.1.1). One sample failed to yield DNA in the 1st and 2nd elution from the column, three samples failed to yield DNA in the 1st elution, however DNA was recovered from the second elution. For the 1st elution off the column the min yield was 13.1 ng/µL with a max of 104.2 ng/µL and a mean of 37.15 ng/µL. The second elution off the column had a min yield was 9 ng/µL with a max of 83.4 ng/µL and a mean of 24.1 ng/µL (Fig 2.4). Elutions were in 100 µL resulting in a minimum yield of 1310 ng for elution 1 and 900 ng for elution 2.

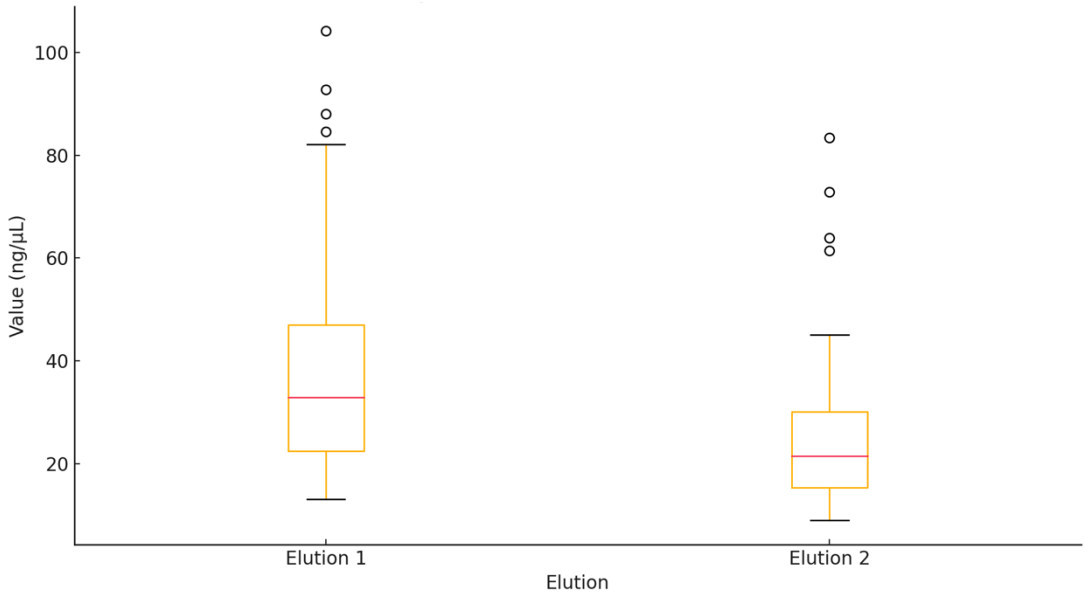

Figure 2.4: Boxplot illustrating the concentration of DNA recovered from the *E. coli* (n = 101) Fire Monkey preparations Elution 1 and Elution 2, measured in ng/µL. Boxplots illustrate the DNA concentration (ng/µL) obtained from 101 UPEC isolates processed using the Fire Monkey HMW DNA extraction *E. coli* SR version 2 protocol.

2.4.2 Developing a Fire Monkey Tecan A200 protocol for *Salmonella*

A key aim of the protocol development for *Salmonella* was to establish a robust method for extracting HMW DNA, while also improving the throughput of the manual steps prior to lysate loading into the A200.

2.4.2.1 Testing

Early testing with clinical *Salmonella* isolates using Protocol 2.4.1.1 encountered two issues. First, the buffer originally used to lyse *E. coli* was ineffective with *Salmonella*. To address this, RevoluGen provided a revised buffer, STET1, which increased the percentage of Triton X-100 from 1.2% to 5%, and included 8% sucrose, 50 mM Tris-HCl, 50 mM EDTA at pH 8. The second issue involved increased clogging of column filters during lysate passage in the A200 run. This was resolved by modifying the A200 flash operations, as detailed in Section 2.4.2.2.

2.4.2.2 Protocol

To begin the preparation, 750 µL of LB was added into each well of a 96-deepwell square-well plate. Using a 10 µL pipette tip, single *Salmonella* colonies were picked and transferred into individual wells, one colony per well, up to a maximum of 96 samples per plate. Once all samples were transferred, the plate was gently swirled for a few seconds to resuspend the cells, and the used tips were discarded. The plate was then sealed with a gas-permeable adhesive film and incubated overnight at 37 °C with shaking at 100 rpm.

After incubation, the plate was placed on ice and centrifuged at 4°C at 4000 rpm using an Eppendorf 5810R centrifuge. This step was critical for efficiently pelleting the cells and greatly facilitated the removal of supernatant. The supernatant was carefully removed by pipette and discarded. Any residual volume of approximately 50 µL was not problematic for subsequent steps. Each well then received 100 µL of STET1 buffer containing 30 mg/mL lysozyme. The STET1 buffer consisted of 8% sucrose, 50 mM Tris-HCl, 50 mM EDTA, pH 8.0, 5% Triton X-100. The mixture was gently pipetted five times to ensure thorough mixing. Notably, STET1 can be prepared in bulk and stored at room temperature, whereas lysozyme was freshly added on the day of use.

The plate was sealed with a standard adhesive plate seal and incubated at 37°C for 10 minutes in a static incubator. Following this, a mixture of LSDNA buffer and proteinase K was added. Specifically, 20 µL of 20 mg/mL proteinase K was added to 300 µL of LSDNA, and 320 µL of this solution was dispensed into each well. The contents were mixed by pipetting five times, and the plate was resealed and incubated at 56°C for 20 minutes in a water bath. The water level was adjusted such that the plate sat at the bottom of the bath without being submerged, and an Eppendorf tube rack was placed on top of the plate to prevent the plate from floating.

Next, 10 µL of 20 mg/mL RNase A was added to each well, followed by gentle pipette mixing and a 5 minute incubation at room temperature. To facilitate DNA precipitation, 350 µL of Binding Solution (BS) was added and mixed five times using a wide-bore pipette tip. Subsequently, 400 µL of 75% isopropanol was added and mixed in the

same manner. The contents of each well were then transferred to a Fire Monkey 96-column plate, which was secured into the Tecan A200 96-column bracket.

Before initiating the automated extraction, buffer levels in the Tecan A200 system were verified. The following volumes per sample were required: 500 µL of WS buffer, 500 µL of 90% ethanol, 200 µL of EB buffer, and at least 500 mL of deionised water (dH_2O). Once the bracket was fixed in position, the Tecan A200 was powered on, and the *Salmonella* program was initiated (Table 2.5). Step one of the *Salmonella* protocol sees a further increase in pressure to ensure all lysate passes through the filter (Fig 2.5).

Table 2.5: Tecan A200 Operations for *Salmonella* Protocol

Step	Operation	Parameter1	Solvent (v/v)	Volume (µL)
1	Flash	salmo load		
2	Wash	500 µL	EtOH:WS (50:50)	500
3	Flash	w1 ec		
4	Wash	500 µL	EtOH:H_2O (90:10)	500
5	Flash	w2 ec		
6	Flash	QIAamp 96 Viral RNA - Drying 30mins		
7	Message	Place Collection Plate		
8	Elute	100 µL	Elution Buffer	100
9	Wait	10 min		
10	Flash	elution ec		
11	Message	Place collection plate		
12	Elute	100 µL	Elution Buffer	100
13	Wait	10 min		
14	Flash	elution 2 ec		
15	Message	Method complete		

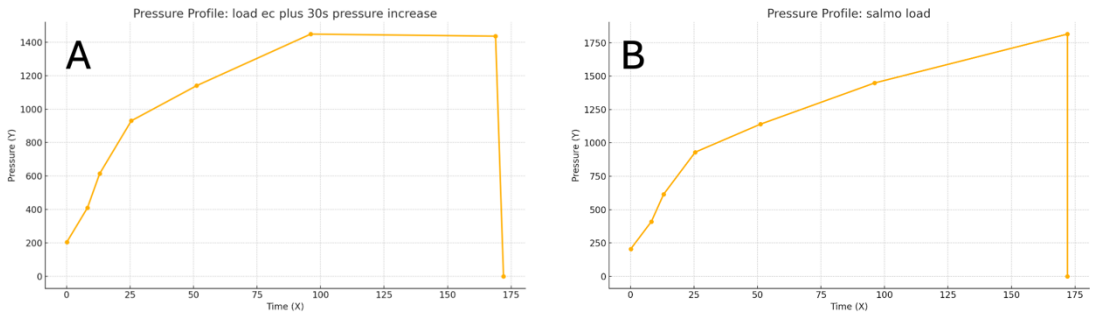

Figure 2.5: Pressure and time setting for Step 1 (lysate loading phase). A, E. coli SR version 2 protocol for the Tecan A200. B, *Salmonella* protocol. Graphs illustrate an increase in pressure and time need to allow complete passage of the lysate through all filters when extracting HMW DNA from *Salmonella*

After Step 10 a pause prompts removal of the columns and bracket to be returned with the addition of a polypropylene fully skirted 96-well collection plate. The run was resumed to allow collection of the first DNA fraction. The same process was repeated using a fresh plate to collect the second fraction. Upon completion of the extraction, both the generator and the Tecan A200 were powered off.

2.4.2.3 Implementation

The protocol was applied to 230 isolates across five independent runs. Elution 1 failed in 13 isolates, while elution 2 failed in four. When elution 2 failed it was in unison with elution 1 failing. For the 1st elution off the column the min yield was 14.4 ng/µL with a max of 134.7 ng/µl and a mean of 51.4 ng/µL. The second elution off the column had a min yield of 2.5 ng/µl with a max of 177.4 ng/µL and a mean of 24.1 ng/µL (Fig 2.6). Elution's were in 100 µl resulting in a minimum yield of 1440ng for elution 1 and 247 ng for elution 2.

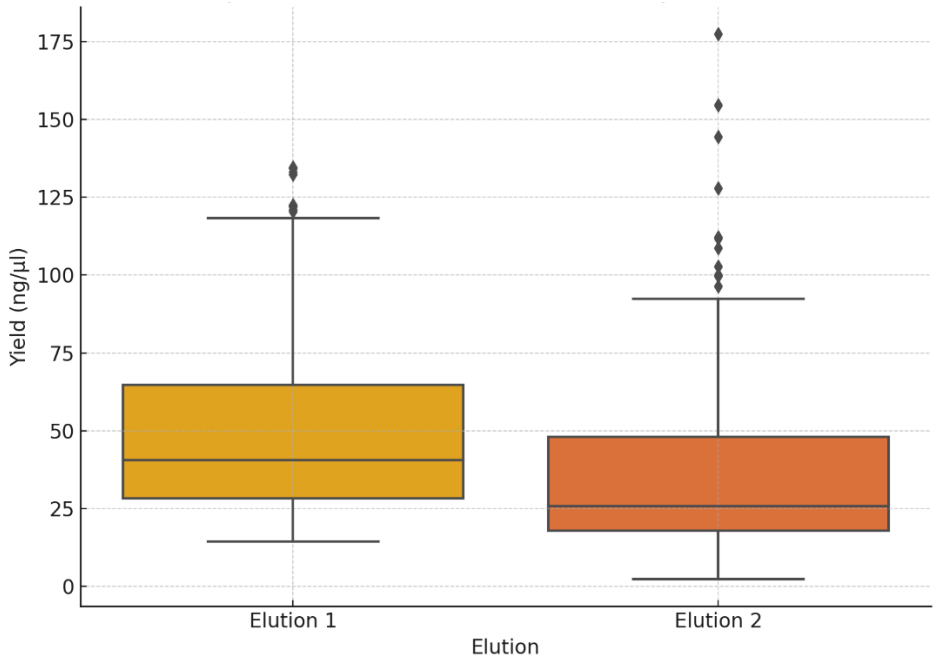

Figure 2.6: Boxplot illustrating the concentration of DNA recovered from *Salmonella* (n = 230) Fire Monkey preparations Elution 1 and Elution 2, measured in ng/µL. Boxplots show the distribution of DNA concentrations (ng/µL) obtained from 230 Salmonella isolates processed using the *Salmonella* Fire Monkey HMW DNA extraction protocol. The protocol was run across five independent batches.

Seven DNA extractions were run on an Agilent TapeStation Genomic DNA ScreenTape to assess the size of the DNA, three are shown in Figure 2.7. The TapeStation struggles with the size of the DNA fragments in these preparations which results in maxing out effect in the results, nonetheless it can be used to see if the DNA is degraded or HMW DNA via DIN and trace peak shoulders. To get a clearer picture on the size of the DNA fragments in these seven samples were sent to RevoluGen where they were analysed using an Agilent Femto Pulse. The Femto Pulse traces show that elution 2 contain a larger average DNA fragment size compared to elution 1. The average size of the DNA in elution 1 was 55-66 kb and in elution 2 55-105 kb (Figure 2.8).

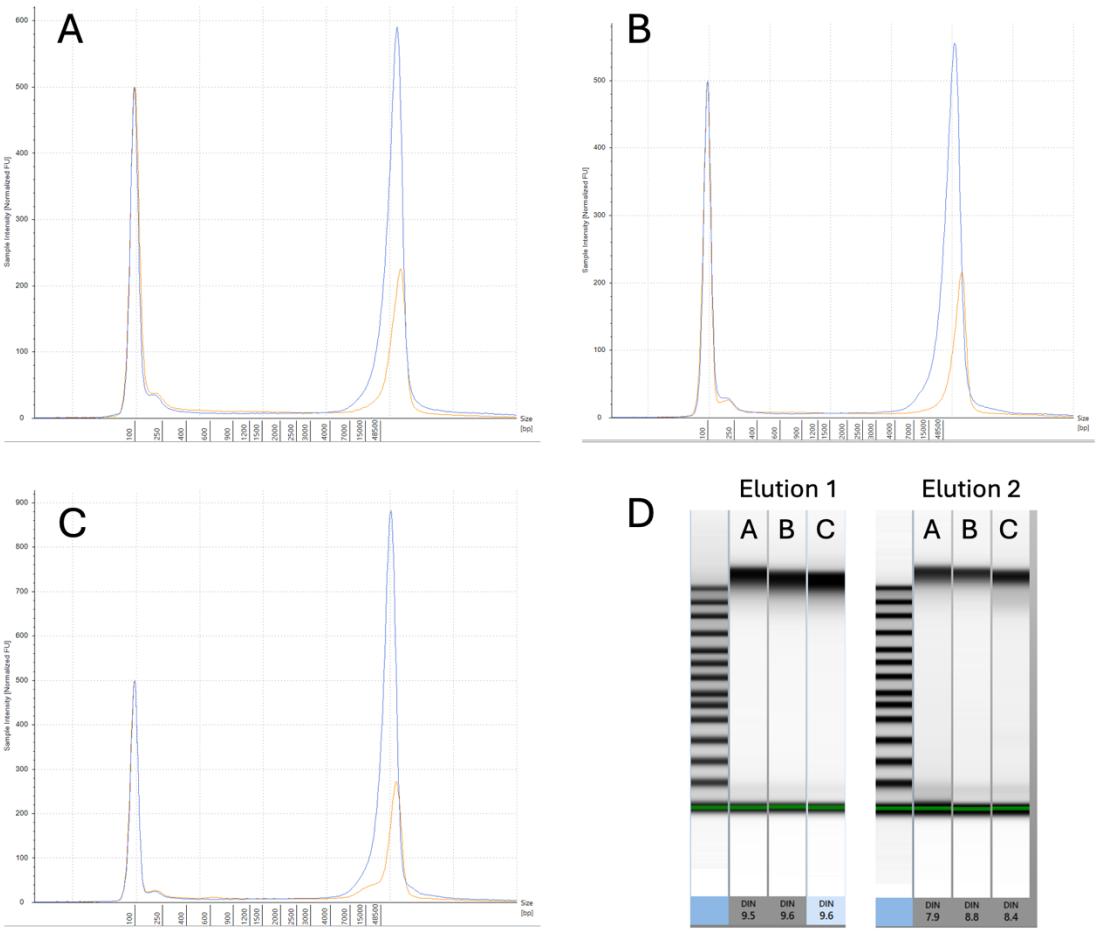

Figure 2.7: TapeStation electrophoresis traces for Fire Monkey extracted *Salmonella* HMW DNA. A-C are individual DNA extractions run on a Genomic DNA ScreenTape, the blue traces represent elution 1 and the orange traces represent elution 2. D Genome DNA ScreenTape shows gel images of the same sample from elution 1 and 2.

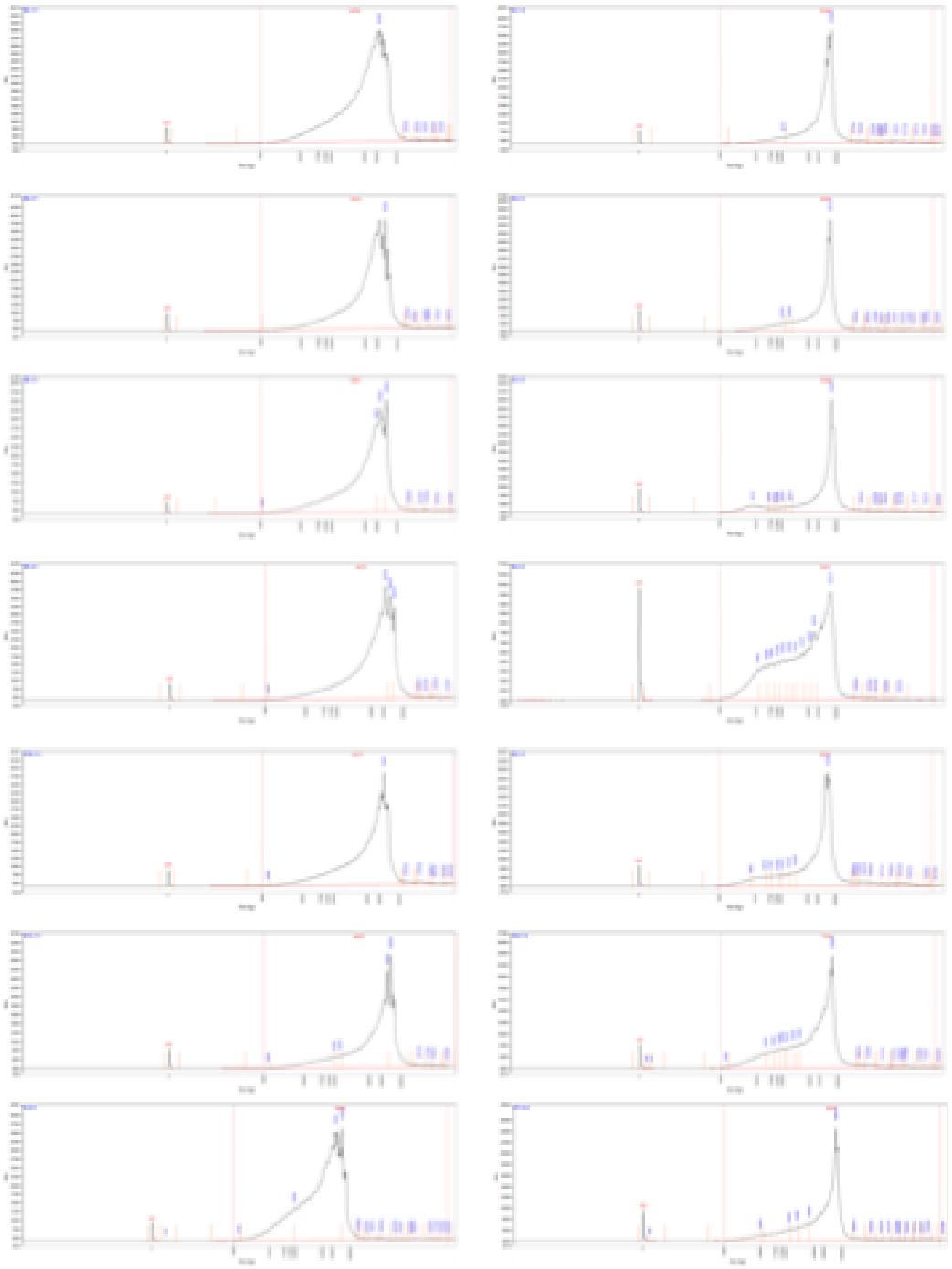

Figure 2.8: Femto Pulse capillary electrophoresis of Fire Monkey *Salmonella* (n=7) HMW DNA extractions using the Tecan A200. Images are paired left and right, left is elution 1 and right is elution 2. Femto Pulse traces are shown for seven Salmonella DNA extractions processed using the Fire Monkey protocol on the Tecan A200 platform. Each pair of images displays Elution 1 (left) and the corresponding Elution 2 (right). Femto Pulse analysis revealed that Elution 2 consistently contained higher average fragment sizes compared with Elution 1. Across the seven samples, Elution 1 produced DNA fragments averaging 55–66 kb, whereas Elution 2 yielded fragments ranging from 55–105 kb. These data provide higher-resolution size profiling than the TapeStation results shown in Figure 2.7 and confirm the presence of HMW DNA suitable for long-read sequencing.

2.4.3 Developing a Fire Monkey Tecan A200 protocol for stool

In this section I present my round-by-round exploration of DNA extraction from stool using Fire Monkey and the Tecan A200.

2.4.3.1 Host depletion

Because human DNA can dominate total nucleic acids in stool, reducing host DNA prior to extraction was necessary to improve bacterial DNA recovery and downstream sequencing efficiency. To prepare stool samples for DNA extraction, I employed a host depletion protocol. Stool was transferred to a 2 mL tube and centrifuged at 16,000 rpm for 2 minutes. After removing the supernatant, the sample's weight was recorded, varying with each test. The standard protocol involved using 200 mg of stool treated with 200 µL of HD buffer, 35 µL of 1% Saponin, and 10 µL of HL-SAN enzyme (ArcticZymes Technologies, Norway). The mixture was vortexed for 30 seconds; occasionally, a pipette tip was used to dislodge the stool pellet, facilitating thorough mixing. The sample was then incubated at 100 rpm for 20 minutes, followed by centrifugation at 13,000 rpm for 5 minutes. After discarding the supernatant, the sample was resuspended in 500 µL of PBS and centrifuged again at 13,000 rpm for 5 minutes. The resulting pellet represented a bacterial cell–enriched fraction suitable for evaluating extraction performance across different lysis strategies.

2.4.3.2 Benchmark: Promega Maxwell RSC Fecal Microbiome DNA kit

Prior work at Quadram had identified the Promega Maxwell RSC Fecal Microbiome DNA kit (Max-RSC) as a promising mid-throughput system for extracting HMW DNA from stool. This system and kit were chosen as the benchmark against which the Fire Monkey protocol was tested. As standard 200 mg of host depleted stool was used as input into the Maxwell RSC Fecal Microbiome DNA kit protocol. To each sample, 1 mL of Lysis Buffer and 40 µL of Proteinase K were added, followed by vortexing for 30 seconds. The tubes were then placed into a heat block at 95°C for 5 minutes, after which they were allowed to cool for 2 minutes on the benchtop. Thorough vortexing for 1 minute was performed before incubating the samples at 56°C for an additional 5 minutes. During this time the cartridges were prepared in accordance with the kit's protocol. Cartridges intended for use were positioned in the deck tray(s). Each

cartridge was securely snapped into place by pressing down, followed by careful removal of the entire seal to ensure all sealing tape and residual adhesive were completely cleared before placement in the instrument. A plunger was then inserted into well #8 of each cartridge, positioned nearest to the Elution Tube. Empty Elution Tubes were placed in the corresponding positions, ensuring their caps were open and facing away from the cartridge positions. Next, 100 µL of Elution Buffer was added to the bottom of each specifically provided 0.5 mL Elution Tube. Subsequently, 300 µL of Binding Buffer was added to well #1 of every cartridge, followed by 20 µL of RNase A to well #3. Following incubation of the sample, the lysate tubes underwent centrifugation at room temperature for 5 minutes at maximum speed (>10,000 × g) to pellet solids. A 300 µL volume of supernatant was transferred into well #1 of the reagent cartridges. The Maxwell device was run using settings Maxwell RSC Fecal Microbiome DNA kit v1.0. This system was selected as a benchmark to evaluate whether the Fire Monkey approach could achieve comparable yields and DNA integrity while providing greater flexibility for automation on the Tecan A200.

2.4.3.3 CTAB extraction

In an initial experiment, I investigated the compatibility of the CTAB lysis buffer from the Maxwell RSC (Max-RSC) kit with the LSDNA buffer and filter columns from the Fire Monkey kit. This experiment was carried out to explore whether the efficient, broad-spectrum lysis properties of CTAB could be combined with the Fire Monkey system's ability to preserve HMW DNA. The Max-RSC kit's lysis method is both rapid and robust, leveraging CTAB for its wide range of activity beyond that of the single digestive enzymes used in the Fire Monkey protocol. Conversely, the LSDNA buffer in the Fire Monkey kit plays a crucial role in maintaining DNA integrity throughout extraction. Here the samples were going through the Max-RSC lysis steps, a 95°C for 5 minutes step followed by a 56°C for 5 minutes step to produce lysate. Once lysate was obtained the Fire Monkey samples were plugged back into the *Salmonella* Fire Monkey protocol (see section 2.4.2.2) for RNase treatment and preparation for binding to the column filter. For the Fire Monkey samples various mixtures of CTAB and LSDNA were used in combination with 30 µL Proteinase K. Five aliquots of stool were used in this experiment, aliquot 1 (271 mg), aliquot 2 (252 mg), aliquot 3 (256 mg), aliquot 4 (270 mg), and aliquot 5 (264 mg) (Table 2.6). Aliquot 1 acted as a control running through the Max-RSC protocol. Aliquot 2 used 1000 µL CTAB + 30 µL Proteinase K in line with the

Max-RSC protocol. Aliquot 3 used 500 µL CTAB with 500 µL LSDNA + 30 µL Proteinase K. Aliquot 4 used 500 µL CTAB + 30 µL Proteinase K with 500 µL LSDNA added after the 95°C for 5 minutes incubation. Finally, Aliquot 5 used 250µL CTAB with 750 µL LSDNA + 30 µL Proteinase K. At this stage, all Fire Monkey extractions were conducted at a higher volume than standard for the *Salmonella* Fire Monkey protocol specifications. This adjustment facilitated extractions using both CTAB and CTAB mixtures, aligning with the volumes specified in the Max-RSC protocol. A 450 µL volume for each aliquot (2-5) was moved to a fresh tube and processed through the Fire Monkey protocol, starting at the RNase treatment stage, followed by addition of BS, and precipitation with 75% isopropanol. These aliquots were then loaded into filter columns and processed using the *E. coli* SR version 2 protocol.

Following extraction, DNA yield and quality were measured to assess the impact of each mixture on recovery efficiency. In the Maxwell extraction (aliquot 1), one cartridge yielded 6.59 ng/µL in a final volume of 100 µL, thus, 271 mg of stool resulted in a total yield of 659 ng. For the experimental samples: aliquot 2, the first elution produced 1.66 ng/µL in approximately 80 µL, but the second elution was too low to measure accurately. From 252 mg of stool, this sample yielded a total of 133 ng. In aliquot 3, the first elution yielded 1.06 ng/µL in about 80 µL, and the second elution was 0.006 ng/µL in approximately 80 µL, resulting in a total yield of 85.3 ng from 256 mg of stool. Aliquot 4 showed 1.05 ng/µL in the first elution and 0.17 ng/µL in the second elution, totalling 97.8 ng from 270 mg of stool. Aliquot 5 yielded no DNA due to filter collapse (Table 2.6). Overall, while CTAB alone produced measurable yields, combining CTAB with LSDNA reduced total recovery, suggesting chemical incompatibility between the two systems and indicating that CTAB-based lysis was unsuitable for direct integration into the Fire Monkey workflow.

Table 2.6: CTAB Extraction Experiment Protocol Variant with Stool Input Weight and Resulting DNA Yield

Aliquot	Condition	Total yield (ng)	Stool weight (mg)
1	Max-RSC protocol control	659	271
2	1000 µL CTAB + 30 µL Proteinase K	133	252
3	500 µL CTAB + 500 µL LSDNA + 30 µL Proteinase K	85.3	256
4	500 µL CTAB + 30 µL Proteinase K, then 500 µL LSDNA after 95°C incubation	97.8	270
5	250 µL CTAB + 750 µL LSDNA + 30 µL Proteinase K	No yield (filter collapse)	264

2.4.3.4 Enzymatic digestion

A more standard approach for the Fire Monkey kit is to use enzymatic digestion. In this experiment I look at the performance of the Fire Monkey kit when using lysozyme treatment and lysozyme plus mutanolysin treatment versus the Max-RSC as a control. Some recommendations from RevoluGen were also tested, these included additional steps: 1). After addition and mixing of BS the samples were rested at room temperature for 5 minutes. In the standard protocol there is no rest time, 2). After the rest the samples were centrifuged at full speed (16,000 x g) for 15 minutes and the samples were then carefully transferred to new tubes. An orange/brown oily looking solution formed in the bottom of tubes was avoided during transfer. Six aliquots of stool were used in this experiment, aliquot 1 (249 mg), aliquot 2 (255 mg), aliquot 3 (242 mg), aliquot 4 (253 mg), aliquot 5 (260 mg), and aliquot 6 (258 mg) (Table 2.7). Aliquot 1 acted as a control running through the Max-RSC protocol. Aliquots 2 and 4 were run throughs of the *Salmonella* Fire Monkey base protocol, aliquot 2 used lysozyme and aliquot 4 used lysozyme plus mutanolysin. Aliquots 3 and 5 were used for run throughs of the extended Fire Monkey protocol, aliquot 3 used lysozyme and aliquot 5 used lysozyme plus mutanolysin. Lysozyme was used at 30 mg/µL in STET buffer and mutanlysin at 25 U in 100 µL STET buffer. Aliquot 6 was run though the Max-RSC protocol with no host depletion, all other aliquots were prepared using the host depletion protocol. These aliquots were then loaded into filter columns and processed using the *E. coli* SR version 2 protocol.

In the Maxwell extraction (aliquot 1), one cartridge yielded 21.0 ng/μL in a final volume of 100 μL, thus, 249 mg of stool resulted in a total yield of 2100 ng. For the experimental samples: aliquot 2, the first elution produced 13.6 ng/μL in approximately 80 μL and the second elution was yielded 1.48 ng/μL in approximately 80 μL. From 255 mg of stool, this sample yielded a total of 1088 ng in the first elution and 118 ng in the second elution. Aliquot 3, the first elution produced 17.9 ng/μL in approximately 80 μL and the second elution was yielded 3.07 ng/μL in approximately 80 μL. From 242 mg of stool, this sample yielded a total of 1430 ng in the first elution and 246 ng in the second elution. Aliquot 4, the first elution produced 17.5 ng/μL in approximately 80 μL and the second elution was yielded 3.13 ng/μL in approximately 80 μL. From 253 mg of stool, this sample yielded a total of 1400 ng in the first elution and 250 ng in the second elution. Aliquot 5, the first elution produced 26.0 ng/μL in approximately 80 μL and the second elution was yielded 3.03 ng/μL in approximately 80 μL. From 260 mg of stool, this sample yielded a total of 2080 ng in the first elution and 242 ng in the second elution. Finally, aliquot 6 produced 31.5 ng/μL in 100 μL from 258 mg of stool, this sample yielded a total of 3150 ng (Table 2.7).

Table 2.7: Enzymatic Digestion Experiment Protocol Variant with Stool Input Weight and Resulting DNA Yield

Aliquot	Condition	Total yield (ng)	Stool weight (mg)	DIN
1	Max-RSC protocol control	2100	249	6.5
2	Fire Monkey base protocol with lysozyme	1206	255	2.3
3	Extended Fire Monkey protocol with lysozyme	1676	242	3.9
4	Fire Monkey base protocol with lysozyme + mutanolysin	1650	253	2.2
5	Extended Fire Monkey protocol with lysozyme + mutanolysin	2322	260	2
6	Max-RSC protocol (no host depletion)	3150	258	6.9

DIN = Agilent TapeStation DNA Integrity Number

The DNA yield for all samples in this experiment was an improvement on the attempt to use CTAB. All enzymatic digestions yielded sufficient DNA for an ONT library preparation protocol with the lowest yielding sample (aliquot 2) yielding 1088 ng. Aliquots 3 and 5 yielded more DNA than aliquots 2 and 4. This was a promising result supporting the additional steps in the protocol aimed at cleaning the lysate before loading into the filter column. The use of lysozyme and mutanolysin together yielded

more DNA than the use of lysozyme. This result supported the use of cocktails of enzymes for a broader bacterial extraction or singular enzymes for a more targeted extraction. The yield in the second Fire Monkey elution was again much lower than first elution. I suspected the amount of DNA going into the filter column was much lower compared to the isolate version of the preparation. Analysis using a TapeStation GenomeTape revealed significant DNA sheering occurring in the Fire Monkey preparation (Figures 2.9 and 2.10). The DNA Integrity Number (DIN) values for the Fire Monkey preparations were much lower than the Max-RSC preparation. All DIN values were lower than a bacterial isolate DNA preparation which should register >8.0. A higher DIN reflects greater intactness of genomic DNA, whereas a lower DIN indicates more degradation. The DIN value is automatically computed using the TapeStation analysis software.

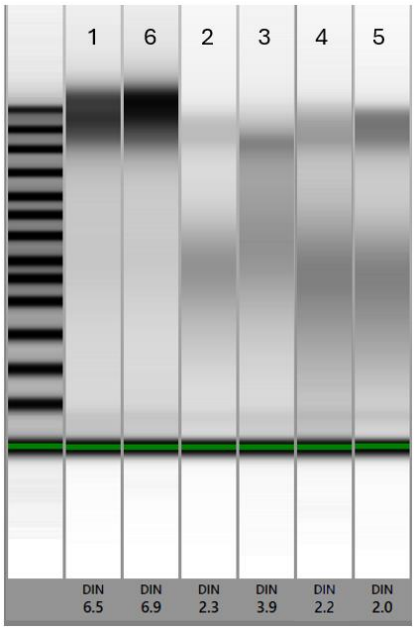

Figure 2.9: TapeStation Genome DNA ScreenTape gel image showing aliquots 1-6 from the enzymatic digestion experiment (n =1). Using elution 1 for the Fire Monkey samples 2-5. Using elution 1 for the Fire Monkey samples 2-5. This Genome DNA ScreenTape gel shows DNA integrity for aliquots 1–6 generated during the enzymatic digestion experiment. Aliquot 1 represents the Max-RSC protocol control, aliquot 2 the Fire Monkey base protocol with lysozyme, aliquot 3 the extended Fire Monkey protocol with lysozyme, aliquot 4 the Fire Monkey base protocol incorporating lysozyme and mutanolysin, aliquot 5 the extended Fire Monkey protocol incorporating lysozyme and mutanolysin, and aliquot 6 the Max-RSC protocol performed without host-depletion. Aliquots 2–5 reflect Fire Monkey preparations using Elution 1. The gel images demonstrate substantial DNA shearing in all Fire Monkey-based preparations compared with both Max-RSC controls, consistent with the lower DIN values observed for these samples and highlighting the sensitivity of HMW DNA integrity to differences in lysate preparation and enzymatic treatment.

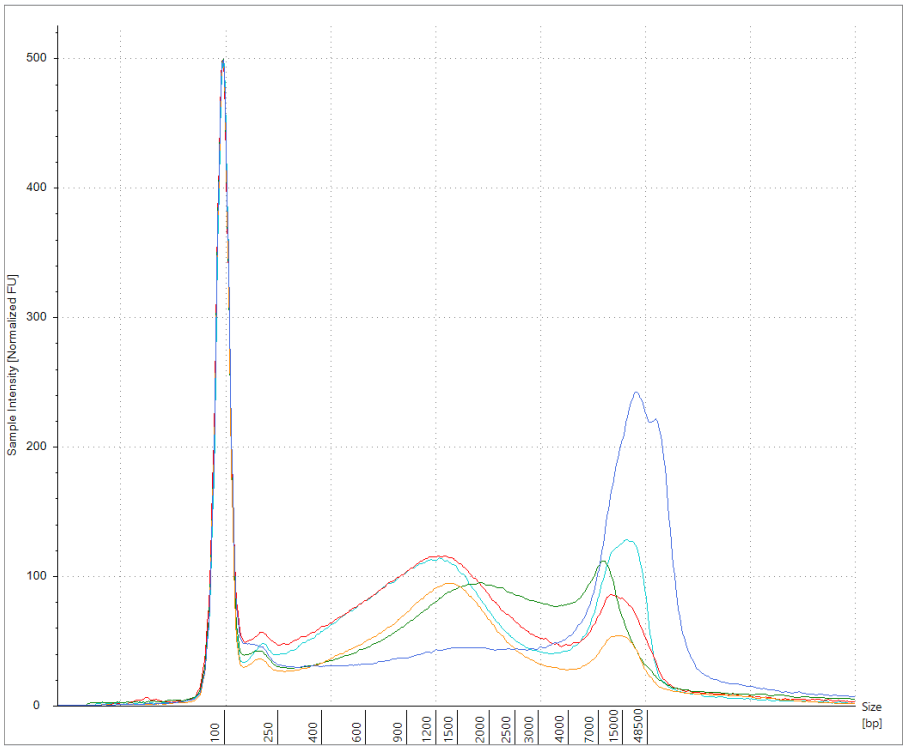

Figure 2.10: Electropherogram from TapeStation Genome ScreenTape for aliquots 1-5 from the enzymatic digestion experiment (n =1). This figure shows Genome DNA ScreenTape electropherograms for aliquots 1–5. Aliquot 1 corresponds to the Max-RSC protocol control (dark blue trace), aliquot 2 to the Fire Monkey base protocol with lysozyme (orange), aliquot 3 to the extended Fire Monkey protocol with lysozyme (green), aliquot 4 to the Fire Monkey base protocol incorporating lysozyme and mutanolysin (red), and aliquot 5 to the extended Fire Monkey protocol incorporating lysozyme and mutanolysin (light blue). All Fire Monkey traces represent Elution 1. The electropherograms confirm extensive DNA shearing across all Fire Monkey conditions relative to the intact high-molecular-weight DNA observed in the Max-RSC control, supporting the DIN findings and illustrating the impact of enzymatic treatment and lysate preparation on DNA fragment length.

The take home message from this experiment was that enzymatic digestion showed promise, but significant DNA shearing was occurring during the preparation as seen in the GenomeTape electropherogram and DIN values. Cleaning the lysate during the preparation improved DNA yield and potentially had a positive effect on the DIN score.

2.4.3.5 Addition of stool washing steps

This section brought me to one of those reflective moments in life where I found myself pondering how I ended up here, meticulously handling and cleaning human stool for a living. In this context, it's crucial to navigate through key inhibitory substances found in

stool, including complex carbohydrates, bile salts, and proteins like mucins and digestive enzymes, alongside challenging components such as humic acids. Adding to these considerations were my concerns about the 5 M salt concentration in the host depletion buffer. In this experiment I looked at washing the stool after host depletion with PBS and warm (55°C) water in an attempt to reduce the salt content. The sample size for this optimisation experiment (n = 6, one aliquot per treatment) was intentionally limited, as the primary aim was exploratory, to identify whether washing steps could mitigate the effects of high salt concentrations from the host depletion buffer without compromising DNA recovery. Each aliquot represented a distinct treatment condition, enabling a direct qualitative comparison of yield, purity, and extraction performance under differing wash regimes. While this single-replicate design was sufficient to identify promising trends and procedural issues, it does not allow for statistical inference or robust quantification of variability between treatments. The weight of the samples was as follows; aliquot 1 (212 mg), aliquot 2 (216 mg), aliquot 3 (219 mg), aliquot 4 (210 mg), aliquot 5 (216 mg), and aliquot 6 (220 mg). Aliquot 1 was Fire Monkey no washes non-host depleted, aliquot 2 was Fire Monkey no washes host depleted, aliquot 3 was wash one time with PBS, aliquot 4 was washed three times with PBS, aliquot 5 was wash one time with warm water, aliquot 6 was washed three times with warm water. The washes consisted of resuspending the stool pellet with a wide bore pipette tip in 1 mL of PBS or warm water. The samples were then centrifuged at 13,000 rpm for 5 minutes and the supernatant was removed. This process was repeated three times for aliquots 4 and 6. During the Fire Monkey DNA extraction process the steps which had a positive effect in section 1.4.3.4 were implemented, those being resting the sample at room temperature for 10 minutes after the addition and mixing with binding solution followed by a centrifuge step, 15 minutes at 16000 x g. These aliquots were then loaded into filter columns and processed using the *E. coli* SR version 2 protocol.

Aliquot 1, the Fire Monkey run with no host depletion steps and no washing, yielded 5.45 ng/µL in 80 µL, a total of 436 ng. Aliquot 2 Fire Monkey run with host depletion and no washing yielded 2.08 ng/µL in 80 µL, a total of 166 ng. Aliquot 3 Fire Monkey run with host depletion with one PBS wash yielded 3.87 ng/µL in 80 µL, a total of 310 ng. Aliquot 4 Fire Monkey run with host depletion with three PBS wash yielded 8.03 ng/µL in 80 µL, a total of 642 ng. Aliquot 5 Fire Monkey run with host depletion with one warm

water wash yielded 5.23 ng/µL in 80 µL, a total of 418 ng, and aliquot 6 Fire Monkey run with host depletion with three warm water wash yielded 14.4 ng/µL in 80 µL, a total of 1150 ng. TapeStation traces of the DNA highlight aliquot 6 as the best performing condition with the highest DIN score (6.1) and a nice peak forming centred on the 15000 bp mark (Table 2.8 & Figures 2.11-2.12).

Table: 2.8 Addition of Stool Washing Experiment Protocol Variant with Stool Input Weight and Resulting DNA Yield Plus DNA Integrity Number (DIN)

Aliquot	Condition	Yield (ng/µL)	Total yield (ng)	Stool weight (mg)	DIN
1	Fire Monkey, no host depletion, no washing	5.45	436	212	2.7
2	Fire Monkey, host depletion, no washing	2.08	166	216	1.8
3	Fire Monkey, host depletion, 1x PBS wash	3.87	310	219	4.9
4	Fire Monkey, host depletion, 3x PBS wash	8.03	642	210	5.6
5	Fire Monkey, host depletion, 1x warm water wash	5.23	418	216	4.8
6	Fire Monkey, host depletion, 3x warm water wash	14.4	1150	220	6.1

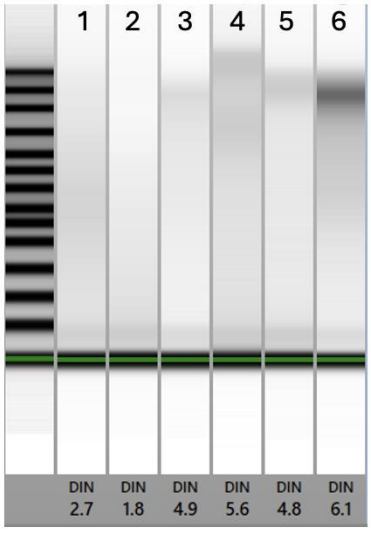

Figure 2.11: TapeStation GenomeTape gel showing DNA from additional washing step testing (n =1). This GenomeTape gel displays the DNA obtained from six aliquots processed under different washing regimes following host depletion. Aliquot 1 represents the Fire Monkey protocol without host depletion or washing; aliquot 2 includes host depletion but no washing; aliquot 3 includes host depletion followed by one PBS wash; aliquot 4 includes host depletion followed by three PBS washes; aliquot 5 includes host depletion followed by one warm-water wash; and aliquot 6 includes host depletion followed by three warm-water washes. All aliquots

were subsequently processed using the Fire Monkey workflow with the optimised resting and high-speed centrifugation steps from Section 1.4.3.4. The gel image illustrates clear qualitative differences between treatments, with aliquot 6 producing visibly HMW DNA and reduced smearing compared with other wash conditions and controls, consistent with its higher DIN value.

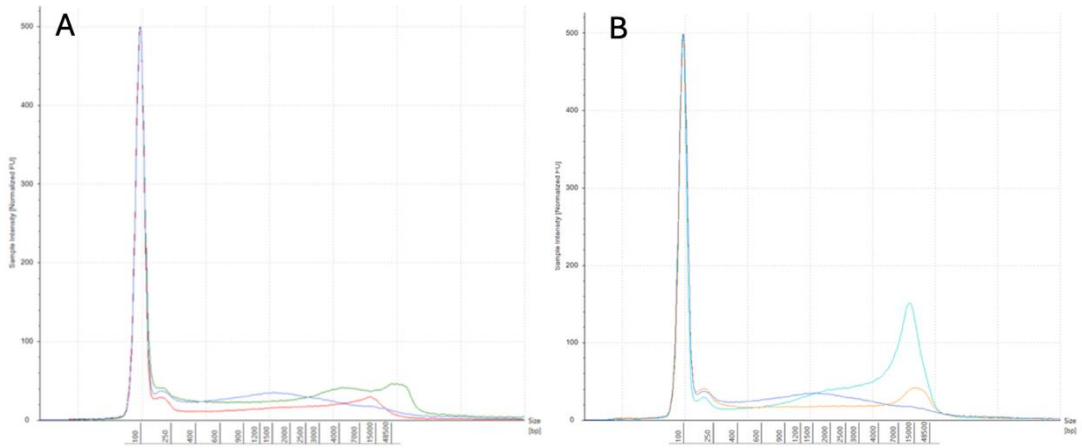

Figure 2.12: TapeStation GenomeTape electropherograms for samples from additional washing step testing (n =1). The electropherograms compare DNA fragment distributions across the washing treatments described in Figure 2.11. Panel A shows the effect of PBS washing, with one PBS wash (red trace) and three PBS washes (green trace) compared against the host-depleted, no-wash control (blue trace). Panel B shows the effect of warm-water washing, with one warm-water wash (orange trace) and 3x warm-water washes (aqua trace) compared to the same host-depleted control (blue). Across both panels, the 3x warm-water wash condition produces the strongest peak and the cleanest electropherogram profile.

The take home message from this experiment was washing the stool pellet three times with warm water after host depletion improved DNA yield and integrity. With each round of washing the supernatant removed was an orange/brown colour suggesting leaching of some compounds of stool origin into the water. This process was not exhausted with three rounds of washing. Washing showed a lot of promise and there were several directions that could be taken for further improvement, namely volume, repetitions, and washing solution.

2.4.3.6 Addition of stool washing steps – Alcohol washes

This one felt a little bonkers at the time, but it was fun to try! The warm water washing had improved the Fire Monkey preparation in terms of DNA yield and DNA integrity. The water wash was primarily aimed at absorbing salts, yet compounds believed to be inhibitory, such as lipids and insoluble proteins, likely persisted in the stool despite the washing process. To target this the host depleted stool pellet was washed in alcohols, ethanol and isopropanol. The stool sample used in this experiment was a diarrhoeal sample that was low yielding when pelleted leading to the use of 80 mg input per preparation versus the standard 200 mg used in previous development experiments. Four aliquots of stool were used in this experiment, aliquot 1 (79.5 mg), aliquot 2 (81.6 mg), aliquot 3 (82.1 mg) and aliquot 4 (80.5 mg). Aliquot 1 was a no wash Fire Monkey control, aliquot 2 was a repeat of the 3x warm water washes, aliquot 3 was ethanol washes, and aliquot 4 was isopropanol washes. The alcohol washes were carried out in a 50:50 solution with PBS. These aliquots were then loaded into filter columns and processed using the *E. coli* SR version 2 protocol.

Aliquot 1, the no wash Fire Monkey control, yielded 0.987 ng/µL in 80 µL, a total of 79.0 ng. Aliquot 2 was a repeat of the 3x warm water washes which yielded 5.91 ng/µL in 80 µL, a total of 473 ng. Aliquot 3 was ethanol washes which yielded 1.23 ng/µL in 80 µL, a total of 98.4 ng, and aliquot 4 was isopropanol washes which yielded 0.770 ng/µL in 80 µL, a total of 61.6 ng. The alcohol washes were not successful with DNA yield close to zero. What was becoming clear was the Fire Monkey performs much better with the 3x warm water washes versus no washes (Table 2.9 & Figure 2.13).

Table 2.9: Addition of Alcohol Washing Steps Experiment Protocol Variant with Stool Input Weight and Resulting DNA Yield Plus DNA Integrity Number (DIN)

Aliquot	Condition	Yield (ng/µL)	Total Yield (ng)	Stool Weight (mg)	DIN
1	Fire Monkey, no wash control	0.987	79.0	79.5	2.3
2	3x warm water wash	5.91	473	81.6	6.7
3	Ethanol wash	1.23	98.4	82.1	2.9
4	Isopropanol wash	0.770	61.6	80.5	1.7

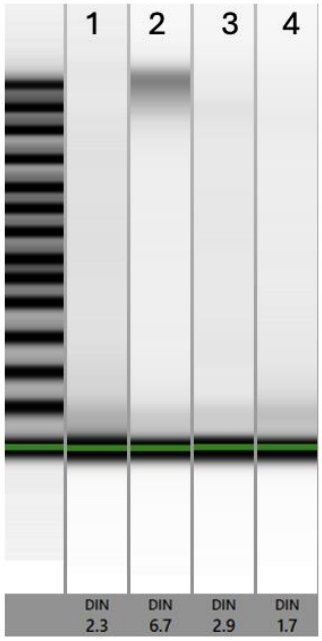

Figure 2.13: TapeStation GenomeTape gel showing DNA from additional washing step including alcohols (n=1). This GenomeTape gel shows the effect of alcohol-based washing steps on DNA integrity compared with water-based washing. Aliquot 1 represents the Fire Monkey protocol with no washing; aliquot 2 includes host depletion followed by three warm-water washes; aliquot 3 includes washing with ethanol; and aliquot 4 includes washing with isopropanol. The gel demonstrates that both ethanol and isopropanol washes were detrimental to DNA quantity and quality, producing pronounced smearing and weaker bands relative to the warm-water wash condition, which yielded the most intact HMW DNA of the treatments tested.

A selection of DNA extractions were again sent to RevoluGen for analysis on the Agilent Femto Pulse. It was interesting to see the difference in profile for the same sample run on a TapeStation versus a Femto Pulse. The distribution of the DNA fragments differed greatly between the two platforms. The TapeStation estimates of the DNA fragmentation profile could be considered overzealous, with a more realistic estimation provided by the Femto Pulse (Figure 2.14).

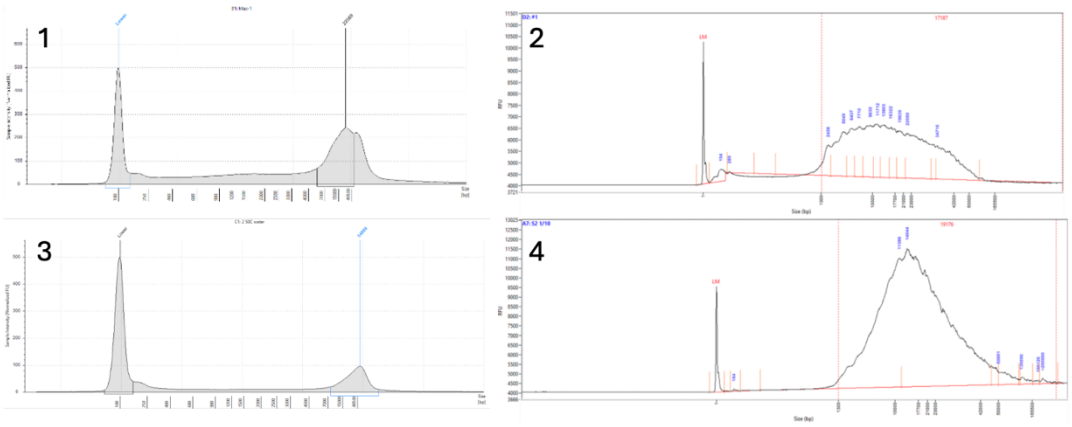

Figure 2.14: Agilent TapeStation and Femto Pulse traces of a Max-RSC stool preparation and a 3x warm water washed Fire Monkey stool DNA preparation (n=1). Trace 1 is a TapeStation trace of the Max-RSC preparation, trace 2 is a Femto Pulse trace of the same Max-RSC preparation. Trace 3 is a TapeStation trace of the Fire Monkey preparation, trace 4 in a Femto Pulse trace of the same Fire Monkey preparation.

The take home message from this experiment was washing the host depleted stool pellet with alcohols was not successful. Washing with warm water once again outperformed preparations with no wash steps. From this point forward I will refer to the Fire Monkey protocol with washing as FM-W; a full version of the protocol can be found in section 2.7.2. The TapeStation was not reliable for accurately sizing this type of DNA, the distribution of fragmentation was underestimated, and the size was overestimated.

2.4.3.7 Addition of Neutrase and increasing reagent volume

Ongoing discussion about the Fire Monkey stool preparation with RevoluGen highlighted the importance of pipette mixing for HMW DNA extraction during the extraction process. The stool had been difficult to move in all experiments described thus far, it stuck to the side/bottom of tubes after being centrifuged, it constantly blocked 1000 µL narrow bore pipette tips and required wide bore tips to resuspend. In the previous experiment, reducing the input from ~200 mg to ~80 mg improved the ease of pipetting the stool and prompted further exploration of reductions down to 50 mg. This adjustment also allowed a greater number of test conditions to be evaluated per stool sample. Neutrase is a broad activity enzyme used to break down proteins of animal and plant origin to peptides. The enzyme had proved useful in

previous work at Quadram exploring microbial communities living on food products (Bloomfield et al., 2023). The Neutrase treatment was applied prior to the host depletion step, 20 µL of Neutrase was added to 1 mL of PBS, this solution was then added to a 50 mg stool pellet. The sample was pipette mixed using a 200 µL wide bore tip and then incubated at 45°C, 200 rpm for 30 minutes. The sample was centrifuged at 5,000 rpm for 5 minutes and the supernatant was removed. The Neutrase treated pellet was used as input for the host depletion protocol and subsequently the Fire Monkey protocol. An alternative strategy to get the stool pellet more suspended in solution was to increase the amount of reagent used during the Fire Monkey preparation. This was implemented to the FM-W protocol, all Fire Monkey protocol steps were carried out in 3x volumes to give the stool a greater volume to resuspend in.

Six aliquots of stool were used in this experiment, aliquot 1 (50 mg), aliquot 2 (54 mg), aliquot 3 (53 mg), aliquot 4 (54 mg), aliquot 5 (50 mg) and aliquot 6 (51 mg). Aliquots 1 and 2 were Fire Monkey without wash steps, aliquots 3 and 4 were FM-W with the Fire Monkey protocol steps carried out at 3x volume, and aliquots 5 and 6 were Neutrase treated Fire Monkey without wash steps.

Aliquots 1 and 2 from the Fire Monkey protocol without washing yielded 13.8 ng/µL and 15.5 ng/µL, respectively. Aliquots 3 and 4 used the FM-W protocol with the Fire Monkey protocol steps carried out at 3x volume yielded 14.5 ng/µL and 11.9 ng/µL, respectively. Aliquots 5 and 6 the Neutrase treated samples yielded 5.11 ng/µL and 7.23 ng/µL, respectively. The Neutrase treatment did not show signs of being beneficial for the preparation with its performance being worse than the Fire Monkey protocol without washing steps. The FM-W protocol with the Fire Monkey protocol steps carried out at 3x volume was the best performing preparation producing a very clean peak on the TapeStation with a DIN score of 8-8.5 (Table 2.10, Figure 2.15 & 2.16).

Table 2.10: Addition of Neutrase and Increasing Reagent Volume Experiment Protocol Variant with Stool Input Weight and Resulting DNA Yield Plus DNA Integrity Number (DIN)

Aliquot	Condition	Yield (ng/µL)	Total yield (ng)	Stool weight (mg)	DIN
1	Fire Monkey, no wash control (FM)	13.8	1100	50	5.7
2	Fire Monkey, no wash control (FM)	15.5	1240	54	6.5
3	FM-W (3x volume), no wash steps	14.5	1160	53	8.0
4	FM-W (3x volume), no wash steps	11.9	952	54	8.5
5	Neutrase treated Fire Monkey, no wash	5.11	409	50	5.7
6	Neutrase treated Fire Monkey, no wash	7.23	578	51	4.6

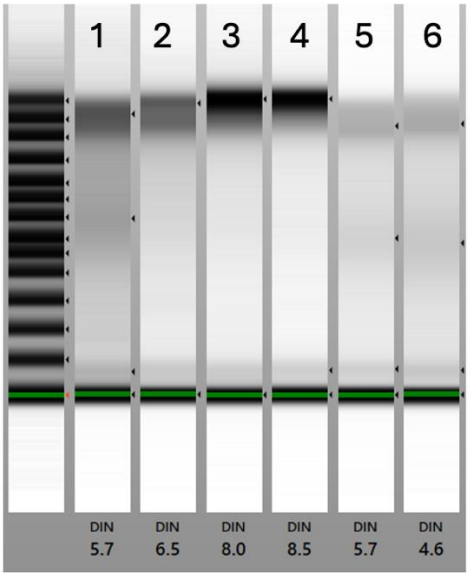

Figure 2.15: TapeStation GenomeTape gel showing DNA from addition of Neutrase and increasing reagent volume testing (n=2). 1 & 2 = Fire Monkey protocol without washing, 3 & 4 FM-W protocol with the Fire Monkey protocol steps carried out at 3x volume, and 5 & 6 Neutrase treated samples.

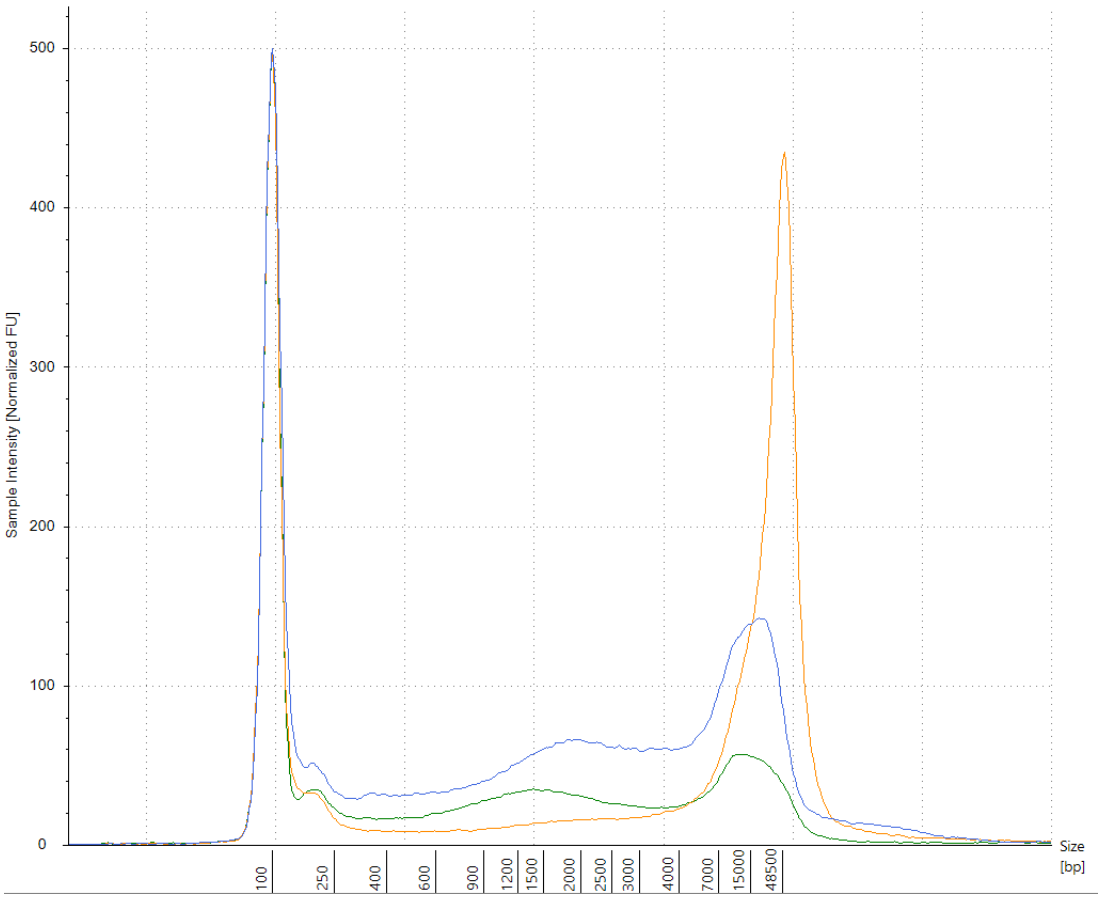

Figure 2.16: The blue traces represent the Fire Monkey Protocol with no wash, the orange represents the FM-W protocol with the Fire Monkey protocol steps carried out at 3x volume, and the green represents the Neutrase treated sample. This electropherogram is showing aliquots 1, 3, and 6, (n=2).

The DNA from aliquots 1-4 was sent to RevoluGen for analysis on the Femto Pulse. The traces show that while the yields (ng/µL) were closely matched, the fragment distribution in the Fire Monkey protocol with 3x warm water washes was more widely distributed and included a larger proportion of DNA fragments 17kb and above (Figure 2.17) The FM-W protocol yielded samples with an average size (bp) 9,480 and 13,644 versus the FM-W at 3x volume protocol yields samples with an average size (bp) 30,789 and 40,952. There was very little DNA in size >=50kb in the FM-W with 0% and 1.6% versus 9.7% and 15.3% for the FM-W at 3x volume protocol.

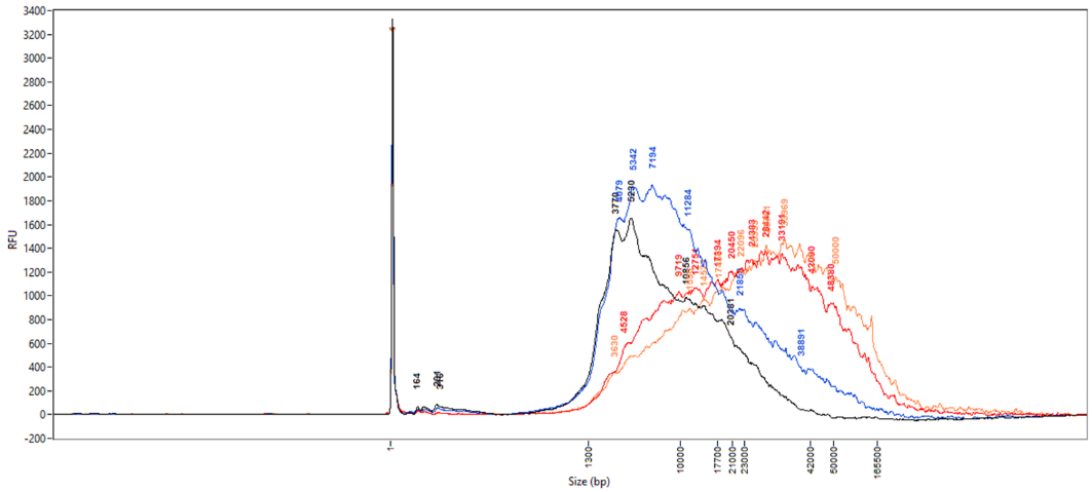

Figure 2.17: Black and blue lines represent the Fire Monkey protocol without washes (n=2) and the red and orange lines represent the FM-W protocol with the Fire Monkey protocol steps carried out at 3x volume (n=2).

The take home message from this experiment was that Neutrase added 1 hour and 20 minutes to the run time of the protocol and did not offer any advantage over warm water washes. Increasing the volume of the reagents during the Fire Monkey protocol has greatly improved the integrity on the DNA as seen in the TapeStation DIN score and in the Femto Pulse traces. The FM-W protocol run with 3x Fire Monkey reagents volumes will be referred to as FM-W-3x, a full version of this protocol can be found in Section 2.7.3.

2.4.3.8 Max-RSC versus FM-W-3x

At this point in the development process the FM-W-3x protocol was comparable with the Max-RSC kit. To test the protocol four stool samples were collected and DNA was extracted using Max-RSC (Protocol in section 1.4.3.2) and FM-W-3x (Protocol in Section 2.7.3). Notably the input for Max-RSC was 200 mg and the input of FM-W-3x was 50 mg. The reduction from 200 mg to 50 mg during Fire Monkey protocol development was necessary to improve DNA integrity and size. The Max-RSC was run using the standard protocol, with the input remaining at 200 mg. The host depletion protocol was run on all samples (Protocol in section 1.4.3.1). The stool IDs for this experiment were 74, 144, 145, and 146. For stool ID 74 Max-RSC yielded 16.5 ng/µL and FM-W-3x yielded 23.0 ng/µL. For stool ID 144 Max-RSC yielded 59.0 ng/µL and FM-W-3x yielded 14.7 ng/µL. For stool ID 145 Max-RSC yielded 58.0 ng/µL and FM-W-3x yielded 33.9 ng/µL. Finally, stool ID 146 Max-RSC yielded 23.5 ng/µL and FM-W-3x

yielded 8.07 ng/µL (Table 2.11). Unfortunately, the TapeStation failed and needed repair at this point in the project, so the reliability of the result provided here is questionable however the preparations all appeared to be of similar size with banding intensities that matched the DNA yields via Qubit assay (Fig. 2.18).

Table 2.11: Max-RSC Versus FM-W-3x Experiment Protocol Variant with Stool Input Weight and Resulting DNA Yield

Stool ID	Protocol	Yield (ng/µL)	Total yield (ng)	Stool weight (mg)
74	Max-RSC	16.5	1650	200
74	FM-W-3x	23.0	1840	50
144	Max-RSC	59.0	5900	200
144	FM-W-3x	14.7	1180	50
145	Max-RSC	58.0	5800	200
145	FM-W-3x	33.9	2710	50
146	Max-RSC	23.5	2350	200
146	FM-W-3x	8.07	645	50

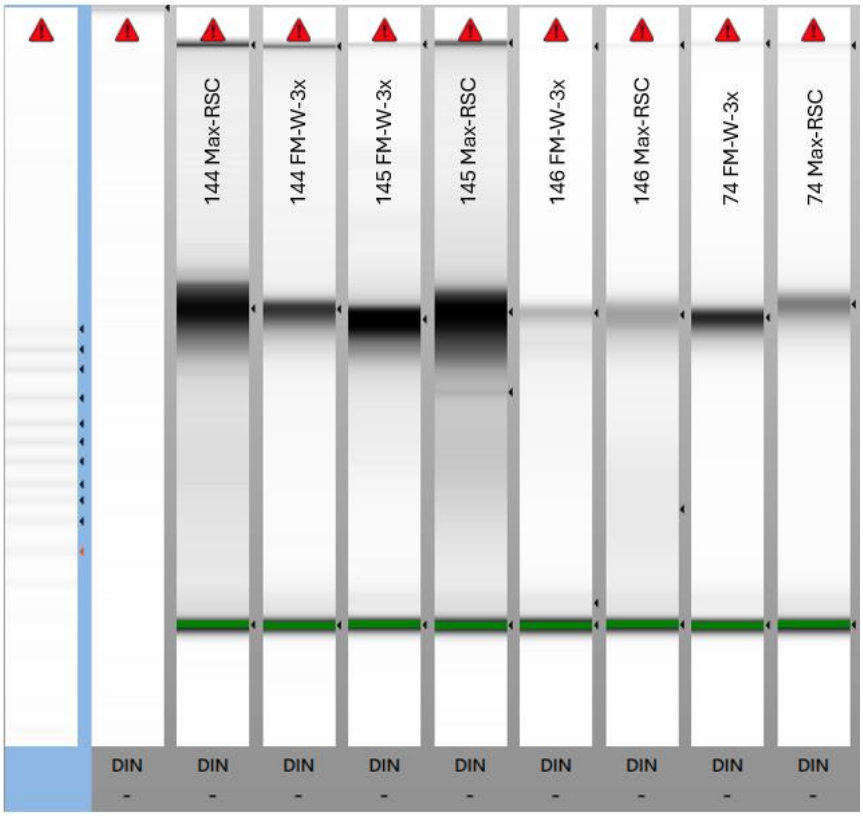

Figure 2.18: TapeStation GenomeTape on DNA from Max-RSC versus FM-W-3x testing. DNA samples are marked on gel lane (n=1).

To further assess DNA purity, Nanodrop measurements were used to obtain 260/280 and 260/230 absorbance ratios. The 260/280 offers a measure of DNA purity with a target score of 1.8 with an acceptable range being 1.7-2.0. A 260/280 at or above 2.0 is generally accepted as being RNA and scores below 1.7 suggest contamination with proteins. The 260/230 offers another measure with 2.0-2.2 being accepted as pure DNA. Below a 260/230 of 2.0 suggest contamination of the DNA with carbohydrates, peptides, or detergents. A 260/230 score above 2.2 suggests RNA contamination or contaminating solvents or detergents, column residue, or faecal derived polysaccharides or certain sugars. All samples in this section were bead cleaned using a 1x SPRI. Neat and bead cleaned samples were analysed using a Nanodrop. Across the sample set the FM-W-3x protocol produced DNA that was less pure than the Max-RSC based on 260/280 and 260/230. Both preparations could produce DNA that was within the target 1.7-2.0 range for the 260/280 and in most cases this ratio could be brought into range using a bead clean. The 260/230 is more problematic for both preparations as a bead clean did not bring the purity of the DNA into the target range (Table 2.12). The take home message from this experiment was that DNA was coming out of the preparations with purity issues and a bead clean was not sufficient to clean the DNA.

Table 2.12: Nanodrop Values for Stool DNA Extracts Pre- and Post-SPRI Bead Clean

Sample ID	Clean status	ng/µL	Total ng	260/280	260/230
144 FM-W-3x	neat	14.0	2120	1.47	0.46
	bead clean	26.0	1270	1.69	1.19
144 Max-RSC	neat	371	24500	1.76	1.30
	bead clean	296	14500	1.76	1.23
145 FM-W-3x	neat	40.0	5480	1.72	0.77
	bead clean	19.0	946	1.67	1.46
145 Max-RSC	neat	125	7750	1.73	1.40
	bead clean	121	5930	1.77	1.34
146 FM-W-3x	neat	18.0	2000	1.49	0.84
	bead clean	14.0	681	1.89	1.00
146 Max-RSC	neat	54.0	4420	1.76	1.26
	bead clean	37.0	1810	1.82	1.39
74 FM-W-3x	neat	25.0	1960	1.58	0.83
	bead clean	43.0	2090	1.84	1.72
74 Max-RSC	neat	37.0	5150	1.58	0.83

	bead clean	19.0	941	1.82	1.12

2.4.3.9 One last look at cleaning

A few final approaches were tested to see if I could improve on the FM-W-3x protocol. Two reagents were selected for their proposed ability to improve DNA purity during the extraction process. InhibitEX is a commercial reagent from Qiagen designed to remove inhibitors from nucleic acid samples during in DNA extraction processes. PVPP is particularly known for its ability to bind and remove polyphenolic compounds and other interfering substances from solutions. Both reagents were added into the FM-W-3x protocol in two positions. The first target was to use the new reagents to clean the stool in place of the warm water and the second target was to add the new reagents to the lysate after treatment with lysozyme. For washing one wash was carried out using the warm water approach followed by a single incubation with 500 µL Inhibitex or 500 µL PVPP or 500 µL of Inhibitex + 500 µL of PVPP. These samples were resuspended using a narrow bore pipette tip and then incubated for 10 minutes at room temperature. The samples were then centrifuged for 5 minutes at 3000 x g the supernatant was removed and the pellet lysed in line with the FM-W-3x protocol. When treating the lysate 800 µL of InhibitEX or 100 µL of PVPP was added to the sample after lysis with STET1 and lysozyme, the samples then followed FM-W-3x. Two more conditions were tested in this experiment, the use of -20°C isopropanol instead of room temperature isopropanol to precipitate the DNA, and finally FM-W-3x was run without host depletion to assess the effect on that step. Input for all conditions was 48-52 mg of stool. All conditions were run in triplicate except the InhibitEX + PVPP wash which was run in duplicate.

Attempts to use the additional reagents as a wash failed. I believe this was due to the action of the compounds and their need to pellet out the contaminates. This caused an issue with the washing procedure as the final product is a stool pellet. In an attempt to pellet the contaminates but retain bacteria in solution, the samples with InhibitEX, PVPP or InhibitEX + PVPP were centrifuged at 3000 x g for 5 minutes and the supernatant was collected. Based on the DNA yield this approach appeared unsuccessful as the bacteria may have pelleted and been lost during the washes. Substituting room temperature 75% isopropanol for -20°C isopropanol also had a negative effect on DNA yield and both Nanodrop ratios. The lysate washes with

InhibitiEX and PVPP showed some promise as the DNA extracts were within range for 260/280, however the 260/230 were well out of target range and the DNA yield was lost in one PVPP lysate wash sample. The no host depletion versus the FM-W-3x protocol suggested there was no detrimental effect on the DNA extract quality when including the host depletion step (Table 2.13).

Table 2.13: Nanodrop Values for Stool DNA Extracts From Final Cleaning Testing

Protocol	Replicate	Nano (ng/µL)	260/280	260/230
FM-W-3x	1.1	29.6	1.87	1.54
	1.2	26.5	1.64	1.58
	1.3	18.9	1.67	3.01
-20 isopropanol	2.1	19.0	0.87	0.34
	2.2	18.5	1.44	1.07
	2.3	24.1	1.78	1.87
InhibitEX lysate	3.1	13.9	1.75	4.64
	3.2	10.1	1.83	-2.32
	3.3	10.4	1.74	-6.43
PVPP lysate	4.1	1.50	2.16	-0.16
	4.2	15.1	1.72	11.42
	4.3	15.0	1.73	5.55
InhibitEX wash	5.1	-2.10	1.32	0.18
	5.2	-3.40	1.47	0.25
	5.3	-2.60	1.62	0.18
PVPP wash	6.1	-1.10	2.53	0.09
	6.2	-1.30	2.47	0.11
	6.3	-1.40	1.26	0.14
FM-W-3x - No host depletion	7.1	29.7	1.79	3.71
	7.2	39.6	1.81	2.28
	7.3	25.2	1.77	4.33
InhibitEX + PVPP wash	8.1	-1.20	1.90	0.10
	8.2	0.40	0.95	-0.04

From the replicates 150 µL DNA and 90 µL SPRI beads were added together to perform a 0.6x SPRI bead clean. The lower ratio SPRI enabled some size selection with fragments smaller than ~600 bp binding less efficiently. Once the samples had been cleaned by SPRI beads the non-host depleted FM-W-3x registered as the purest samples followed by the FM-W-3x protocol. The 260/230 remained a slight issue (Table 2.14).

Table 2.14: Nanodrop Values for SPRI Bead Cleaned DNA From Final Cleaning Testing

Protocol	Nano (ng/µL)	260/280	260/230
FM-W-3x	50.0	1.84	1.69
-20 isopropanol	28.0	1.70	1.72
InhibitEX lysate	25.7	1.71	1.53
PVPP lysate	23.8	1.72	1.62
InhibitEX wash	0.70	0.97	-4.10
PVPP wash	4.00	1.14	1.31
FM-W-3x - No host depletion	67.9	1.84	1.86
InhibitEX + PVPP wash	1.50	0.72	2.04

To conclude I have not found a cleaning solution more effective than warm water. The input to DNA extraction was a DNA pellet making it hard to use reagents that chelate contaminates to be removed by pelleting. Cleaning the lysate shows some potential but would need some testing to optimise the technical aspects of those protocol steps.

2.4.3.10 Final adjustments

I have been burnt a couple of times running stool DNA on Nanopore flowcells so I was determined to chase down pure textbook DNA samples before my final sequencing attempt. In the last experiment (Section 2.4.3.9) I had noticed during the centrifuge steps of the lysate cleaning steps a substantial pellet of debris was formed. So, the final adjustment was to add a gentle 3000 x g centrifuge step to the lysate after RNase treatment and before addition of binding solution. I had some excess stool, so a few extra conditions were included in this final experiment. I was intrigued by the potential benefits of sorbitol, which is thought to support osmotic balance and stabilise cellular membranes during lysis. By regulating osmotic pressure, sorbitol helps to prevent DNA degradation and maintain the integrity of the DNA molecules extracted from cells. I created buffer STET2 which in comparison to STET1 substitutes the 8% sucrose is for 9% sorbitol. I also trialled washing purely with 9% sorbitol instead of warm water. A single stool sample was used in this experiment, which was the same stool sample as in 1.4.3.9; all aliquots were 50-53 mg, two aliquots were run for the four conditions.
In this experiment the STET1 lysis buffer performed better. The DNA neat out of the preparations using STET1 had a higher DNA yield plus 260/280 and 260/230 ratios

closer to the pure DNA range. The lysate spin step greatly improved the 260/230 ratios directly out of the preparations. A SPRI bead clean on the DNA resulted in the FM-W-3x protocol with the lysate spin yielding pure DNA with both 260/230 and 260/230 ratios being in range. This scenario was also true for the version of the protocol run with 9% sorbitol washing (Table 2.15)

Table 2.15: Nanodrop Values Pre- and Post-SPRI Bead Clean for Final Adjustment Testing

		Fire Monkey Preps			0.6x SPRI Clean		
Protocol	**Replicate**	**ng/µL**	**260/280**	**260/230**	**ng/µL**	**260/280**	**260/230**
FM-W-3x / Lysate spin	A-1	44.8	1.71	2.57	74.6	1.85	2.01
	A-2	40.6	1.62	2.05			
Water wash / STET2 / Lysate spin	B-1	18.5	1.58	4.34	50.1	1.8	1.74
	B-2	35.2	1.67	2.21			
Sorbitol wash / STET1 / Lysate spin	C-1	40.8	1.70	2.76	69.6	1.84	1.92
	C-2	32.1	1.71	3.07			
Sorbitol wash / STET2 / Lysate	D-1	26.3	1.58	2.49	44.1	1.83	1.65
	D-2	22.7	1.60	3.51			

The take home message from this experiment was that I now had a Fire Monkey protocol that yields µg amounts of pure DNA. The lysis buffer STET1 performed better than STET2 the sorbitol variant. For washing the stool pellet after host depletion 9% sorbitol worked as well as warm water. For now, I will continue with warm water. In the next section (2.4.3.11) the final stool preparation is described. To see how the Fire Monkey Stool DNA protocol performed in sequencing two stools versus the Max-RSC protocol see Chapter 5.

2.4.3.11 A Tecan A200 program for Fire Monkey Stool DNA extraction

An A200 stool protocol was developed during the preparation phase. Due to the volume of lysate generated during the FM-W-3x preparation, the robot needed to be loaded in three instalments. Step 1 was run twice independently to load the first two volumes of lysate onto the column. On the third and final lysate load, the complete A200 stool protocol was executed. The Stool A200 protocol was built from the original *E. coli* protocol (Table 2.16). The lysate was not as DNA laden as the isolate DNA preparations, so the pressure and time was reduced on the lysate load flash

operations and also the pressure and time were reduced for step 3's flash operation (Figure 2.19).

Table 2.16: Tecan A200 Stool Protocol

Step	Operation	Parameter1	Solvent
1	Flash	load stool	
2	Wash	500 µL	EtOH:H_2O 50:50 (v/v)
3	Flash	w1 stool	
4	Wash	500 µL	EtOH:H_2O 90:10 (v/v)
5	Flash	w2 ec	
6	Flash	QIAamp 96 Viral RNA - Drying 30 min	
7	Message Only	Place Collection Plate	
8	Elute	100 µL	H_2O Tris (EB)
9	Wait	10 min	
10	Flash	elution ec	
11	Message Only		

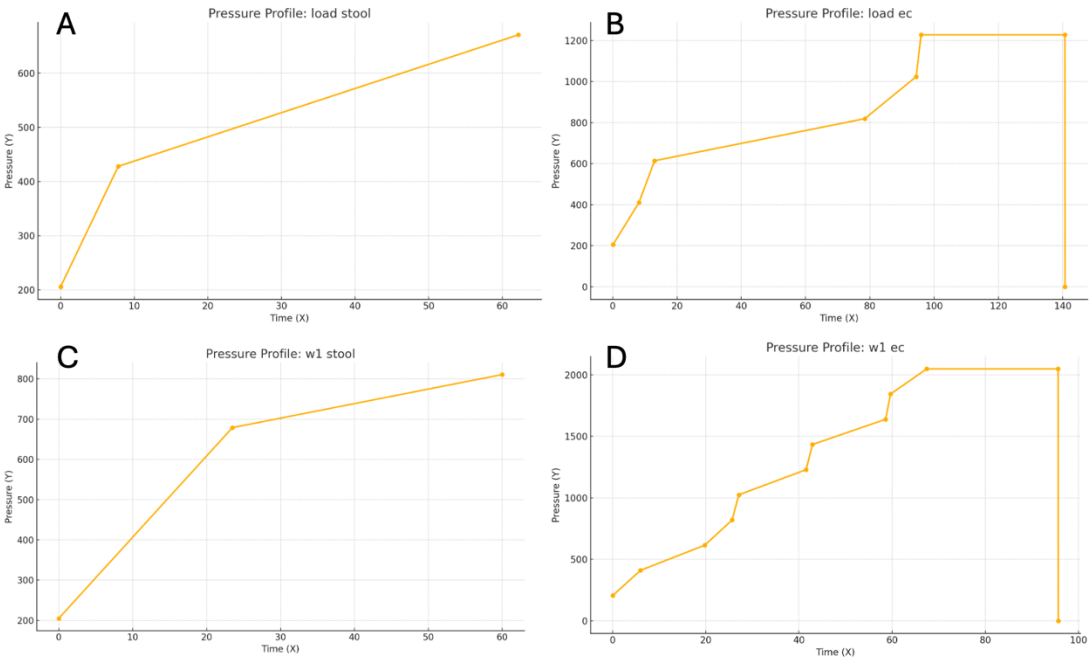

Figure 2.19: Changes made to Tecan A200 pressure profiles for the Stool A200 protocol. A = Step 1 lysate load phase stool protocol, B = Step 1 lysate load phase *E. coli* protocol, C = Step 3 first filter column wash stool protocol, and D = Step 3 first filter column wash *E. coli* protocol.

2.4.3.12 Fire Monkey Stool DNA extraction protocol

DNA extraction: Host depletion (HD)

1. Using a sterile 10 µL loop or a 200 µL wide bore pipette tip, inoculate raw stool into labelled 2 mL round bottom sterile Eppendorf tubes.

Optional: If the stool is too watery, centrifuge diarrheal sample (Bristol scale 6-7) into a 'solid' pellet as a starting sample for HD treatment. Use diarrheal pellet as original sample. Measure out 50 mg.

2. Add 200 µL of HD-buffer to each stool sample tube.
3. Add 5 µL of HL-SAN enzyme to the stool sample with HD-buffer.
4. Add 17.5 µL of 1% saponin.
5. Vortex sample for 30 seconds. If stool is stuck as pellet in tube use a pipette tip to dislodge, return to vortex until pellet is partly resuspended.
6. Incubate sample at 37 °C for 20 minutes on an Eppendorf shaking heat block set to 1000 rpm.
7. Vortex each HD sample tube until mixed.
8. Centrifuge at 18,000 rcf for 3 minutes to pellet HD sample.
9. Gently remove the supernatant.

Washing

1. Add 2 mL of 50 °C sterile water to the pellet.
2. Vortex for 30 seconds.
3. Centrifuge at 18,000 rcf for 3 minutes to pellet HD sample.
4. Gently remove the supernatant.
5. Resuspend in 2 mL of 50 °C dH_2O.
6. Centrifuge at 18,000 rcf for 3 minutes to pellet HD sample.
7. Gently remove the supernatant.
8. Resuspend in 2 mL of 50 °C dH_2O.
9. Centrifuge at 18,000 rcf for 3 minutes to pellet HD sample.
10. Gently remove the supernatant.
11. The HD pellet is considered the input stool sample for Fire Monkey.

Fire Monkey Stool DNA extraction

1. Use a 2 mLtube.
2. Add 300 µL of STET1 (30 mg/mL lysozyme).

3. Resuspend with wide bore tip 1000µl tip (get stool pellet moving)'
4. Continue resuspension with narrow bore 1000 µL tip (x10).
5. Incubate at 37 °C for 10 minutes.
6. Add 900 µl LSDNA and 60 µl (20 mg/mL Proteinase K).
7. Resuspend with wide bore tip 1000 µl tip (get stool pellet moving).
8. Continue resuspension with narrow bore 1000 µL tip (x10).
9. Incubate at 56 °C for 20 minutes.
10. Add 3 µl RNase A (100 µg/µL in H_2O).
11. Incubate at room temperature for 5 minutes.
12. Centrifuge at 3,000 rcf for 5 minutes.
13. Split lysate to three eppendorfs.
14. Add 350 µl of BS to each tube.
15. Resuspend with narrow bore tip 1000 µL tip (x10).
16. Incubate at room temperature for 20 minutes.
17. Centrifuge at 18,000 rcf for 20 minutes
18. Transfer supernatant to fresh tubes, avoid debris. Often yellow/brown oily solution at bottom of tube
19. Add 400 µl of 75 % isopropanol and vortex.
20. Load tube 1 of 3 to A200 plate column, run stool load only protocol.
21. Load tube 2 of 3 to A200 plate column, run stool load only protocol.
22. Load tube 3 of 3 to A200 plate column, run stool protocol.
23. Elution set for 1 x 100 µl (yield in 2nd elution very low, so stopped collecting)

2.5 Discussion

2.5.1 HMW Stool DNA Extraction with Fire Monkey

Extracting HMW DNA from stool is inherently challenging due to the complex and "dirty" composition of faecal material (Reuter & Zaheer, 2016). Stool contains numerous PCR-inhibitory substances including complex polysaccharides, bile salts, urea, glycolipids, heme derivatives, and even residual diet components which can co-purify with DNA and inhibit enzymatic reactions (McGaughey et al., 2019; Paulos et al., 2016). Effective protocols must therefore remove or neutralise these inhibitors early in the process. Additionally, stool harbours a diverse microbiota with tough cell walls (especially Gram-positive bacteria with thick peptidoglycan layers) that are resistant to lysis (McGaughey et al., 2019). If lysis is incomplete, DNA yield will be biased toward easily lysed cells (e.g., Gram-negatives and host cells), skewing the microbial profile (Isokääntä et al., 2024; Roopnarain et al., 2017; Stinson et al., 2019). However, if one is screening for a specific Gram-negative pathogen (e.g., *Salmonella* or *Campylobacter*) then biased lysis can become an enrichment process. Achieving comprehensive lysis often requires intense mechanical disruption (e.g., bead-beating) which can fragment DNA. This presents a trade-off. Methods like bead-beating ensure high yield and representation of microbes, but they shear DNA, limiting the fragment size obtained. Since HMW DNA (≤50 kb) is a primary goal for long-read sequencing, mechanical shear must be minimised to preserve integrity (Maghini et al., 2021; Trigodet et al., 2022).

Another challenge is host DNA contamination. Human epithelial cells and leukocytes are present in stool albeit usually a minor fraction of total DNA, often <10% of reads in healthy samples (Pereira-Marques et al., 2019). In this project (Chapter 4) I show that human reads can dominate the stool DNA extraction with several cases resulting in over 96% of the sequencing reads being of human origin. For microbiome-focused applications, excessive host DNA is undesirable as it reduces the proportion of microbial reads and can mask low-abundance taxa. Moreover, human cells are typically easier to lyse (due to less robust cell membranes), so extraction protocols that are gentle on bacterial cells may disproportionately release host genomic DNA (Bloomfield et al., 2023). The saponin-HL-SAN based host depletion protocol was established before this project started and I have adopted it to aid with the removal of

human DNA from stool samples during this project (T. Charalampous et al., 2019; B. Djeghout et al., 2024).

DNA degradation and shearing are constant concerns when isolating HMW DNA. Faecal samples may contain nucleases or harsh chemicals that can degrade DNA if not quickly inactivated (Reuter & Zaheer, 2016). Standard silica spin column methods often involve multiple binding and wash steps that, if not optimised, may shear HMW DNA through excessive pipetting or exposure to air-water interfaces (Chen et al., 2020; Lever et al., 2015). The Fire Monkey kit was specifically designed to mitigate some of these issues by incorporating a built-in size-selection during extraction, thereby avoiding extra post-extraction handling that could break long DNA. Maintaining long strand length is critical because longer DNA fragments improve long-read assembly and structural variant resolution (Warburton & Sebra, 2023). Thus, an ideal microbiome stool DNA extraction protocol for HMW DNA must balance aggressive lysis to maximise yield from all microbes with gentle handling to preserve strand length, all while removing PCR inhibitors and limiting host DNA carryover. The plug and play nature of the lysis step of the Fire Monkey protocol lends itself well to both targeted DNA extraction and broader microbial extraction. This was show early in the protocol development with the use of lysozyme and mutanolysin yielding more DNA than lysozyme by itself. For targeting specific bacteria for extraction from a complex substrate with a community of species it is tantalising to think with a combination of AI and engineering, enzymes could be developed to specifically target and lyse a narrower range of bacteria greatly improving direct sequencing applications for clinical diagnostics.

Adapting the Fire Monkey HMW DNA extraction protocol (originally validated on cultured cells and mammalian tissues) to stool required iterative optimisation to overcome the above challenges. Early attempts with the unmodified kit on stool samples revealed issues with DNA yield, DNA fragment size and purity, prompting a series of protocol modifications focused on lysis efficiency and inhibitor removal. Key adjustments included enhancing the wash steps and increasing reagent volumes. Additional wash steps were incorporated to more thoroughly rinse away inhibitors and contaminants from the stool before lysis. In the optimised protocol, the host depleted stool pellet is washed three times (FM-W-3x) with water warmed to 50°C. The warm

wash helps to dissolve residual salts and/or bile acids and reduces sample viscosity, thus preventing carryover of these inhibitors. By increasing the number and volume of washes, the modified Fire Monkey protocol achieved higher A260/230 purity ratios indicating fewer organic contaminants.

Stool samples are highly heterogeneous and often viscous, so the volumes of lysis and binding buffers were scaled up to ensure adequate mixing and contact with all particulates. The standard Fire Monkey protocol was designed for relatively clean cell suspensions; by tripling the lysis buffer volume and proportionally increasing downstream reagent volumes, the stool samples could be fully homogenised in solution. This dilution should have helped reduce local concentrations of inhibitors and improved DNA adsorption to the silica matrix by providing a larger volume for chaotropic salts to denature proteins and for DNA to bind. A larger volume also means a portion of the lysate (containing the bulk of inhibitors and debris) could be sacrificed if needed, a strategy akin to that used by some commercial kits. For example, the Maxwell RSC Fecal Microbiome kit lyses stool in a large volume, pellets debris and then only loads 300 µL of the cleared supernatant onto the extraction cartridge, leaving behind inhibitor-laden debris. Similarly, my Fire Monkey adaptations discard pellets of insoluble matter after centrifuge spins at the cost of some DNA, this helps protect the purity and integrity of the remaining HMW DNA.

Together, these modifications transformed the Fire Monkey protocol into a more robust method for stool. Each change was observation and data driven, DNA yield (by fluorometry) and purity (Nanodrop 260/280 and 260/230 ratios) were measured, and the fragment length distribution was checked (TapeStation or Femto Pulse). Over multiple cycles of optimisation, I observed improvements in both yield and integrity leading to a final protocol which can deliver pure DNA based on Nanodrop ratios.

Implementing the above HMW DNA extraction protocol on a Tecan A200 robotic platform conferred significant advantages in consistency and throughput but also introduced new practical considerations. On the benefit side, automating the protocol on the Tecan ensured that each sample was processed with identical timing, mixing, and incubation conditions, thereby improving reproducibility. Automated liquid handling minimises user-to-user variation and reduces the risk of human error in

pipetting, which is particularly valuable for protocols requiring careful layering of reagents or gentle handling of HMW DNA. Indeed, automation of DNA extraction has been shown to produce comparable (or improved) purity and yield versus manual methods, while greatly reducing hands-on time (McGaughey et al., 2019). In our context, the Tecan A200 can handle up to 96 samples per run performing steps post DNA precipitation including DNA collection, in column washing, drying and elution. Whilst this is not full automation it does streamline the DNA extraction process by avoiding the manual labour of repeated filter column-based washes in a centrifuge. A prior comparison of automated vs. manual stool DNA prep demonstrated that an instrument like the Maxwell RSC could extract DNA in ~30 minutes per batch, compared to 1.5–2 hours manually, without compromising downstream sequencing quality (McGaughey et al., 2019). The Fire Monkey protocol is still a long process, without including host depletion the Fire Monkey steps up until the robot is loaded take 115 minutes for stool samples when processed at low numbers (max 20). The Tecan A200 protocol takes 50 minutes, resulting in a full run time of 205 minutes (3 hours 25 minutes). This is the major drawback with the Fire Monkey column-based process as the additional steps to ensure DNA purity have greatly increased the time to complete the protocol.

Stool lysates are not as homogeneous or predictable as the cultured cell or blood sample preparations that many robotic workflows are designed for. The complexity and variability of the lysate meant that clogs and pipetting issues were initial hurdles. For example, early in development I found that pipette tips would often clog with particulate matter when aspirating crude stool lysate. I mitigated this first by using wide bore tips and next by incorporated larger reagent volumes to enable resuspension with narrow bore tips. This experience aligns with general recommendations for automated nucleic acid extraction, samples with a large amount of debris can clog filtration devices or overwhelm bead-based clearing methods if not pre-cleared (Promega, n.d.). The Tecan A200 was equipped with a vacuum filtration module to handle the spin-column binding and washing steps, but very viscous samples still presented challenges in flowing through the silica membranes, sometimes causing slow or incomplete filtration. Warming the lysates and wash buffer during the extraction process helped to reduce viscosity and pipetting with combinations of 1000 µl wide bore, 200 µl wide bore, and 100 µl narrow bore tips.

2.5.2 HMW bacterial cell DNA Extraction with Fire Monkey

Throughout this project the Fire Monkey bacterial cell HMW DNA extraction using the Tecan A200 has been robust yielding DNA concentrations suitable for completing long- and short-read sequencing projects from a single bacterial culture. For *E.coli* and *Salmonella* the size of DNA fragments obtained in both Fire Monkey elution steps was in the ≤50-100 kb range and could be considered HMW DNA. DNA for *E.coli* and *Salmonella* resulting from preparations in this chapter have been successfully sequencing using ONT and Illumina. Results for *Salmonella* can be found in Chapter 3 of this thesis and in published work (Rudder et al., 2025). Results for *E. coli* can be found in work published by our group (Carter et al., 2023). The Tecan A200 proved very easy to use and is intuitive to edit profiles to increase pressure and/or time to ensure the lysate had appropriate conditions to pass through the filter columns. As the Tecan A200 system is semi-automated a 96-well format the front-end to the extraction process was developed to enable high-throughput HMW DNA extraction from bacterial cells. The front of the protocol including stages of cell growth, lysis, and preparing the DNA for binding to the filter column. The Fire Monkey process transferred without issue from the single tube format to the 96-well plate format. The benefits to performing growth and the front-end of the Fire Monkey process in a 96-well plate format is time and throughput. I was able to process 96 samples in the same timespan as 24 samples in single tubes. There are some risks involved with the transfer to 96-well plates which are mainly concerning cross contamination of samples. It is important to maintain a low rotation during growth so that the cultures in the plate are not thrown against the plate seal risking cross contamination of wells. It is also important to use a multi-channel pipette throughout the process to avoid missing cells and care needs to be taken to avoid overflow of wells when adding isopropanol as at this stage of the preparations the volume is close to the maximum volume a well can hold. Finally, having transferred the stages before the use of the Tecan A200 into a 96-well format this opens up opportunities to automate these steps.

While manual adaptation of the Fire Monkey process to a 96-well format enabled high-throughput extraction, it also introduced potential risks of cross-contamination during liquid handling steps. These risks primarily arise from manual pipetting, culture

agitation, and the high fill volumes required during precipitation. In future iterations, integrating these front-end stages with a robotic liquid handling system would substantially reduce contamination risk by standardising pipetting accuracy, minimising manual intervention, and ensuring consistent plate sealing and mixing conditions. Automation would therefore not only improve reproducibility and precision but also further safeguard sample integrity across large-scale extractions.

2.6 Conclusion

The Fire Monkey system has proven itself to be robust for DNA extraction for the bacteria tested and for human stool. The plug and play nature of the lysis is a core strength opening up a wide range of applications from single isolate HMW DNA extraction to broad range microbial DNA extraction from complex sample types to targeted DNA extraction from complex sample types. All stages of the Fire Monkey DNA preparation transfer to 96-well format opening up opportunities to fully automate the process. The main weakness with Fire Monkey is that it is not a fast process, especially when running the stool protocol.

3 Genomic diversity of non-typhoidal *Salmonella* found within patients suffering from gastroenteritis in Norfolk, UK

Chapter contributions: Bilal Djeghout assisted with sample collection.

The work presented in this chapter has been published as:
Rudder, S. J., Djeghout, B., Elumogo, N., Janecko, N., & Langridge, G. C. (2025). Genomic diversity of non-typhoidal *Salmonella* found in patients suffering from gastroenteritis in Norfolk, UK. Microbial Genomics, 11(8), 001468. https://doi.org/10.1099/mgen.0.001468

3.1 Introduction

Salmonella is a prominent public health pathogen, causing a spectrum of disorders such as gastroenteritis, enteric fever, and invasive non-typhoidal salmonellosis (Galán-Relaño et al., 2023; Langridge et al., 2012; Marchello et al., 2020). Deciphering the *Salmonella* genome is fundamental for understand this complex genus which has over 2,600 serovars that differ in epidemiological significance. Accurate genomic characterisation is central to tracing outbreaks and shaping public health responses (Chattaway et al., 2023).

High-throughput, short-read WGS technologies have enabled a transition from biochemical based typing methods to analysis of DNA sequences (Chattaway et al., 2023). Numerous public health agencies around the world use DNA sequencing as their main method for surveillance and outbreak studies as it offers highly accurate genome level information (Brown et al., 2019; Chattaway et al., 2023; Deng et al., 2016; W. Li et al., 2021; Meumann et al., 2022). For standard WGS procedures, one colony from a culture plate is usually chosen as the starting material for DNA extraction and sequencing (Ford et al., 2018; Köser et al., 2012; Kwong et al., 2015). This method gives an accurate picture of a single *Salmonella* genome, but it doesn't take into account the potential diversity that may exist within a single patient (Holt et al., 2009; Raghuram et

al., 2023). This conventional approach leaves a critical gap in our understanding of within-patient *Salmonella* diversity.

Recent studies have identified genome-level diversity within a single-host infection for various human pathogens, including *Burkholderia dolosa (Lieberman et al., 2014)*, *Campylobacter* (Djeghout et al., 2022), *Clostridium difficile* (Eyre et al., 2013), *Helicobacter pylori* (Wilkinson et al., 2022), *Mycobacterium tuberculosis* (Liu et al., 2015), *Staphylococcus aureus* (Raghuram et al., 2023), and *Streptococcus pneumoniae* (Tonkin-Hill et al., 2022). If a patient is infected with multiple strains, STs or a population containing significant SNPs, our ability to effectively conduct surveillance and accurately reconstruct transmission chains from a single colony is compromised.

Recent advances in sequencing, especially the combination of long- and short- read technologies, enable bacterial genomes to be examined in greater detail from structural complete genome assemblies (Wick et al., 2023). By combining the complementary strengths of long and short reads, namely the ability of long reads to resolve complex genome structures with the high accuracy of short reads, hybrid genomes can provide a powerful opportunity to gain insights into genetic diversity, AMR mechanisms, and overall genome architecture from a single assembly (Bouras et al., 2024; Khezri et al., 2021; Waters et al., 2025).

This study investigates genome-level diversity among *Salmonella* isolates recovered from individual patients' stool specimens in Norfolk, UK, using hybrid genome sequencing. By exploring the strengths and limitations of hybrid assemblies, the work evaluates their role in resolving structural features and detecting genomic variations that influence pathogen behaviour. The findings highlight the complexities of genomic analyses, the importance of capturing intra-sample diversity, and the implications for epidemiological investigations and outbreak detection. This research underscores the need for optimised sequencing approaches to ensure accurate and comprehensive genomic insights in both clinical and research settings.

The number of stool specimens analysed in this study (n = 8) was determined by the considerable scale and complexity of the experimental design. Each sample required the isolation of up to 20 individual *Salmonella* colonies, extraction of high-quality DNA, and sequencing using both long- and short-read platforms to enable hybrid genome

construction and comparison. This process involved extensive laboratory work and substantial computational analysis, including multiple genome assemblies, polishing steps, and variant calling. The sample size therefore represented a practical compromise between experimental depth and available resources, allowing for detailed within-host investigation while maintaining feasibility within the project's timeframe and budget.

3.2 Aims and objectives

The work outlined in this chapter aimed to:

- Recover up to 20 *Salmonella* isolates from an individual patient's stool sample
- Leverage hybrid genome assemblies to explore genome-level diversity among isolates from a single patient's stool specimen, to include:
    - Identification of serovar and sequence type
    - Identification of antimicrobial determinants
    - Identification of genome structure
    - Analysis of single nucleotide polymorphisms
- Provide knowledge base for use of single colonies in bacterial diagnostic laboratory sequencing

3.3 Materials and methods

3.3.1 Stool collection

Stool specimens surplus to requirements were collected from the National Health Service (NHS) Eastern Pathology Alliance (EPA) laboratory, Norwich, Norfolk, United Kingdom (UK) between March 2020 and August 2022. Three samples were collected before the start of this project in 2020, and five samples were collected during the project. All samples were marked *Salmonella* spp. positive at the EPA, as determined by a PCR-based culture independent testing panel (Gastro Panel 2, EntericBio, Serosep, United Kingdom). Aliquots of up to 20 mL were transferred triple contained to the Quadram Institute Bioscience (QIB) where they were split and stored as up to 1 mL aliquots raw and as a 50:50 mix with *Brucella* Broth supplemented with 17.5 %

glycerol. These aliquots were transferred to the University of East Anglia (UEA) Biorepository where they were stored at -80 °C. Stool specimens were stored until a serovar was confirmed by the UK Health Security Agency (UKHSA); this was a safety measure put in place to avoid inadvertent cultivation of a Hazard Group 3 *Salmonella* species.

3.3.2 Bacterial isolation

Plastic loops were used to transfer ~10 µL of stool to bi-plates containing Xylose Lysine Deoxycholate (XLD) agar (Oxoid, UK) and Brilliance™ *Salmonella* agar (BSA; Oxoid, UK). The quadrant streak method was applied to obtain single colonies. Plates were incubated at 37 °C for 16 hours. The selective properties of the media were used to identify putative *Salmonella* colonies. Colonies with black centres, observed on XLD agar where the surrounding media remained pink or red, were selected. Purple colonies were selected from BSA. Selected colonies were streaked using the quadrant streak method to MacConkey media (Oxoid, UK) and incubated at 37 °C for 16 hours. Colonies that were circular and remained colourless or pale were selected, as opposed to colonies that caused the media to become pink. A continuous streaking approach was used to propagate putative *Salmonella* colonies onto individual Tryptic Soy Agar (TSA; EO Labs, UK) plates for a final sterility check. These plates were incubated at 37 °C for 16 hours. Using a loop, a significant portion of the bacteria from each TSA plate was collected and stored in Brucella broth supplemented with 17.5 % glycerol at -80 °C.

3.3.3 DNA extraction

Bacterial Culture Preparation: single bacterial colonies were inoculated into 500–1000µL Lysogeny Broth (LB) in a 96-deepwell plate (square well plate). A 10 µL pipette tip was used to pick each colony, which was then introduced into individual wells. The plate was gently swirled to mix. The plate was sealed with a gas-permeable adhesive seal and incubated overnight at 37 °C in an incubator shaker set at 100 rpm for 16–18 hours. Centrifugation and Cell Pellet Preparation: following incubation, the plate was placed on ice and centrifuged at 4 °C and 4000 rpm using an Eppendorf

5810R centrifuge. This step facilitated efficient cell pelleting, which simplified the subsequent removal of the supernatant. The supernatant was carefully removed using a pipette, leaving ~50 µL of residual media to avoid disrupting the pellet. Cell Lysis: each well received 100 µL of STET1 buffer (8% sucrose, 50 mM Tris-HCl, 50 mM EDTA, pH 8.0, 5 % Triton X-100) containing 30 mg/mL lysozyme. The solution was mixed by pipetting up and down five times. The STET1 buffer was prepared in bulk and stored at room temperature, while lysozyme was freshly added on the day of use. The plate was sealed with a standard adhesive plate seal and incubated at 37 °C for 10 minutes in a static incubator. Proteinase K Treatment: a mixture of 20 µL of 20 mg/mL proteinase K and 300 µL LSDNA buffer was prepared and added to each well (final volume: 320 µL per well). The contents were mixed by pipetting five times, and the plate was resealed with an adhesive plate seal. The plate was incubated at 56 °C for 20 minutes in a water bath, ensuring the plate rested on the bottom without being submerged by using an Eppendorf tube rack to prevent floating. RNase A Treatment: after incubation, 10 µL of 20 mg/mL RNase A was added to each well, mixed by pipetting, and incubated at room temperature for 5 minutes. Subsequently, 350 µL BS buffer and 400 µL isopropanol (75 %) were sequentially added, with each step involving mixing five times using a wide-bore pipette. Purification Using Fire Monkey 96-Column Plate: samples were transferred to a Fire Monkey 96-column plate. The column plate was then mounted in a Tecan A200 96-column bracket. The Tecan A200 was prepared for operation, ensuring sufficient volumes of the following buffers were available: WS buffer (500 µL per sample), 90 % ethanol (500 µL per sample), elution buffer (EB; 200 µL per sample), and deionised water (at least 500 mL). The generator and the Tecan A200 were powered on, and the *Salmonella* program initiated. At the program's first pause, the column plate and bracket were removed, placed atop a polypropylene fully skirted 96-well plate, and returned to the Tecan A200 to proceed with the first elution fraction. This process was repeated with a fresh plate to collect the second fraction.

3.3.4 DNA quantification

3.3.4.1 Single tube assay

The Qubit dsDNA BR Assay Kit (Q32853, Thermo Fisher, UK) was used as follows: 199 µL of Qubit™ dsDNA BR buffer and 1 µL of Qubit dsDNA BR Reagent were

combined to prepare a master mix of the appropriate volume. For standards, 190 µL of the master mix was mixed with 10 µL of the Qubit™ dsDNA BR Standards supplied with the kit. For samples, 198 µL of the master mix was mixed with 2 µL of DNA. Each sample was vortexed for 10 seconds and allowed to rest for at least 2 minutes before being measured using a Qubit 3.0 Fluorometer. All standards and samples were quantified using Qubit assay tubes (Q32856, Thermo Fisher, UK). During Oxford Nanopore Technologies (ONT) library preparations (1.2.7.1) 1 µL of DNA library was used with 199 µL of master mix.

3.3.4.2 Plate assay

The Quant-iT dsDNA Assay Kit (Q33130, Thermo Fisher, UK) was used according to the manufacturer's instructions, 199 µL of Quant-iT dsDNA BR buffer and 1 µL of Quant-iT dsDNA BR reagent were combined to prepare a master mix of the appropriate volume. For standards, 190 µL of the master mix was mixed with 10 µL of the λ dsDNA BR standards (0, 5, 10, 20, 40, 60, 80, and 100 ng/µL) supplied with the kit. For samples, 198 µL of the master mix was mixed with 2 µL of DNA. All standard and samples were added to a CytoOne flat bottom, non-treated 96-well plate (CC7672-7696, Starlab, Germany). The plate was gently vortexed, briefly centrifuged and allowed to rest for at least 2 minutes. Readings were taken using a Promega GloMax Discover System (Promega, USA).

3.3.4.3 DNA sizing

DNA integrity and size were estimated using the Genomic DNA ScreenTape analysis (5067-5365 & 5067-5366, Agilent Technologies, USA) on an Agilent TapeStation. DNA sizing was performed using the Agilent Genomic DNA ScreenTape assay according to the manufacturer's protocol. Each sample was prepared by mixing 1 µL of genomic DNA with 10 µL of Genomic DNA Sample Buffer in a PCR tube. For each assay, 1 µL of Genomic DNA Ladder was mixed with 10 µL of Genomic DNA Sample Buffer. All samples were gently vortexed and briefly centrifuged prior to analysis.

3.3.5 DNA cleaning & concentrating

DNA cleaning and concentration were performed using AMPure XP beads (A63881, Beckman Coulter, USA) following an in-house protocol based on the manufacturer's guidelines, a 1:1 ratio of DNA sample to beads was mixed in a 1.5 mL Eppendorf tube and incubated at room temperature for 5 minutes. The tube was then placed on a magnetic rack, allowing the beads to migrate toward the magnet (~2 minutes). Once the supernatant cleared, it was carefully removed. The beads were washed twice with 500 µL of 70 % ethanol, with each wash being gently pipetted over the beads and then removed. Any residual ethanol was removed by pipette, and the tube was air-dried for 30 seconds. After removing the tube from the magnetic rack, elution buffer was added to resuspend the beads by flicking the tube. The sample was incubated at room temperature for 5 minutes before being returned to the magnet, and the eluted DNA was transferred to a fresh Eppendorf tube. The volume of elution buffer varied based upon the level of concentration required.

3.3.6 Bacterial isolate short-read sequencing

DNA normalised to 10 ng/µL was submitted to QIB sequencing for library preparation and sequencing. Miniaturised Illumina DNA Prep kit reactions of 0.5 µL of Tagmentation buffer (TB1), 0.5 µL bead-linked transposomes (BLT), 4 µL molecular grade water, and 2 µL DNA at 10 ng/µL were prepared for each sample. The tagmentation mix was heated for 15 minutes at 55°C in a thermocycler. The 7 µL tagmentation mix was added to the following PCR master mix; 10 µL KAPA 2G Fast Hot Start Ready Mix (KK5601, Merck, UK), 2 µL molecular grade water, 1 µL 10 µM primer mix containing both P7 and P5 Illumina barcodes. The following PCR cycles were run: 72°C for 3 minutes, 95°C for 1 minutes, 14 cycles of 95 °C for 10 seconds, 55 °C for 20 seconds, and 72 °C for 3 minutes. Libraries were quantified by the QIB sequencing facility using Promega QuantiFluor dsDNA System (E2670, Promega, UK) in a Promega GloMax Discover Microplate Reader. After equal-molar pooling of samples the final pool was double Solid Phase Reversible Immobilization (SPRI) size selected between 0.5X and 0.7X bead volumes using sample purification beads supplied with the Illumina DNA Prep kit (Cat. No. 20025519, 20025520, 20018704, and 20018705,

Illumina, USA). Final library quantification and sizing was by Promega QuantiFluor dsDNA System using a Qubit 3.0 instrument and by D5000 ScreenTape (5067-5579, Agilent Technologies, USA) using the Agilent TapeStation 4200. The final pool was run at a concentration of 1.5 pM on an Illumina NextSeq500 instrument using a 300 cycle Mid Output Flowcell (FC-404-2003, Illumina, USA) following the Illumina denaturation and loading recommendations with 1 % PhiX (FC-110-3001, Illumina, USA).

3.3.7 Bacterial isolate long-read sequencing

3.3.7.1 Miniaturised ONT LSK109 library preparation

Long-read sequencing was performed on an ONT MinION using the SQK-LSK109 kit with the EXP-NBD196 barcoding expansion, accommodating up to 48 samples per run. A miniaturised preparation was carried out as follows: ~175 ng of DNA (~14 ng/µL) was used as input. Within a PCR plate, 12 µL of DNA was combined with 0.875 µL Ultra™ II End Repair/dA-Tailing Buffer, 0.375 µL Ultra™ II End Repair/dA-Tailing Mix (E7646, New England Biosciences, UK), 0.875 µL NEBNext FFPE DNA Repair Buffer, and 0.375 µL NEBNext FFPE DNA Repair Mix (M6630L, New England Biosciences, UK), in a total volume of 15 µL. The mixture was incubated in a thermal cycler at 20 °C for 5 minutes, followed by 65 °C for 5 minutes.

For barcoding, 3.75 µL (~87.5 ng) of the end-prepped mixture was combined with 1.25 µL of a native barcode (one unique barcode per sample), 1 µL Blunt/TA Ligase Master Mix (M0367, New England Biosciences, UK), and 4 µL 5x Quick Ligase Reaction Buffer (B6058S, New England Biosciences, UK). The mixture was gently pipette-mixed and incubated in a thermal cycler at 20°C for 120 minutes, followed by 65 °C for 20 minutes. Barcoded libraries were pooled by combining 10 µL of each sample. A 0.6x AMPure XP bead cleanup was performed, and the sample was eluted in 35 µL of nuclease-free water. DNA concentration was assessed using a Qubit assay as described in Section 1.2.4.1 to ensure sufficient material for library preparation.

To attach sequencing adapters, 30 µL of the cleaned pooled library was combined with 5 µL of Adapter Mix II, 10 µL of 5x Quick Ligase Reaction Buffer, and 5 µL of Quick Ligase (M2200S, New England Biosciences, UK). The mixture was flick-mixed and incubated at room temperature for 20 minutes. Next, 30 µL of AMPure XP beads were added to the adapter-ligated library, flick-mixed, and incubated at room temperature for 10 minutes before placing the tube on a magnet. Once the supernatant was removed, 250 µL of Short Fragment Buffer was added to the tube. The sample was removed from the magnet, and the beads were resuspended in the solution before returning the tube to the magnet. This washing step was repeated. After removing the supernatant, the sample was taken off the magnet and resuspended in 15 µL of ONT Elution Buffer. The DNA concentration was checked as described in 1.2.4.1 using 1 µL of sample.

Critical control steps in this miniaturised ONT LSK109 library preparation include ensuring accurate DNA input quantity and integrity, as high-molecular-weight DNA is essential for achieving optimal read lengths and sequencing yield. Precise temperature control during end-repair and dA-tailing reactions is critical for complete enzymatic activity, while correct barcode ligation and strict one-barcode-per-sample handling prevent cross-contamination and misassignment. The 0.6x AMPure XP bead cleanup step must be carefully executed to retain HMW fragments while removing smaller ones, and adapter ligation requires accurate reagent volumes and incubation conditions to maximise sequencing efficiency. Finally, thorough washing with Short Fragment Buffer and accurate quantification of the eluted DNA ensure purity and sufficient concentration for successful flow cell loading.

3.3.7.2 MinION loading

Buffers FLT and FB were held on ice until thawed, 30 µL of FLT was added to 1170 µL of FB and then pipette mixed to create the flush buffer. In this chapter the flow cells used were version R9.4.1. Once a flowcell check was complete 800 µL of the flush buffer was added to the flowcell via the priming port. The process of adding the flush buffer to the flow cell was completed slowly and with caution to ensure no bubbles were introduced into the flowcell. After 5 mins the SpotON sample port was opened and a

further 200 µL of flush buffer was added through the priming port. At this point the flow cell is ready for sample loading.

The sample was prepared for loading as follows: 12 µL of DNA sample was mixed with 37.5 µL SQB and 25.5 µL LB in a fresh Eppendorf tube. The sample was loaded into the SpotON sample port. Flow cells were run for 72 hours.

3.3.7.3 Base-calling

ONT MinKNOW software (v4.0.5) was used to collect sequencing data. Base-calling was performed locally, alongside de-multiplexing and barcode trimming using ONT Guppy (v5.0.11).

3.3.8 Genome assembly

3.3.8.1 Short-read assembly

Short reads from QIB sequencing were uploaded to QIB's data cloud utilising the Integrated Rapid Infectious Disease Analysis (IRIDA) platform by QIB's core informatics team. Paired-end short-read files were imported into Galaxy, a bioinformatic workflow platform hosted by the Norwich Research Park (NRP). Reads were filtered to remove Illumina adaptor sequences and low quality reads with fastp (Galaxy v0.19.5 (Chen et al., 2018)) using default settings, phred quality 15, a limit of 40% for unqualified bases, and a limit of 5 Ns per read, Shovill (Galaxy v1.1.0 (Seemann, 2017)) was used to assemble reads using SPAdes (Bankevich et al., 2012), the Shovill Galaxy v1.1.0 wrapper uses a SPAdes version ≥ 3.14.

3.3.8.2 Long-read assembly

Fastq files were uploaded to Galaxy and filtered using Filtlong (Galaxy v0.2.0 (Wick & Menzel, 2019)) with settings Min. length = 1000, and Min. mean quality = 50. Filtered reads were assembled with Flye (Galaxy v2.5 (Lin et al., 2016)) mode = Nanopore raw, with an estimated genome size set to 4.8 m. The assembly fasta was passed to

medaka (Galaxy v0.11.5) along with the filtered reads for polishing using model r941_min_high_g303. The medaka consensus fasta was passed to racon (Galaxy v1.0.11) along with the filtered reads for two rounds of polishing. During the assembly process checkM (Galaxy v1.0.11(Parks et al., 2015)) was applied to monitor genome completeness and contamination scores after each step.

3.3.8.3 Hybrid assembly

To address challenges encountered with single nucleotide polymorphism (SNP) calling from hybrid genomes in this work, various assembly strategies were explored in this chapter, ultimately leading to the decision to use short reads for SNP calling in the final analysis.

3.3.8.3.1 Hybrid 1. LR-Pilon

A long-read assembly polished with short reads using Pilon. Short reads were filtered using fastp (Galaxy v0.19.5) with default settings, which perform quality trimming (Phred < 15), removal of adapter sequences, filtering of reads shorter than 15 bp, and automatic correction of paired-end read overlap and base errors. Filtered short reads were mapped to long-read assembles using minimap2 (Galaxy v2.12 (Li, 2018)) with setting -Hk19 creating a bam file. The long-read assembles and bam files were passed to Pilon (Galaxy v1.20.1(Walker et al., 2014)) for round one of short-read polishing with min depth setting = 0.2, default base quality = 15, and kmer size = 47. A second bam file was created by mapping short reads to the round one polished fasta files. The second bam file and round one polished fasta files were passed to Pilon to complete the second round of polishing. The same setting were used in both rounds of polishing. This fasta file was assessed using checkM (Galaxy v1.0.11).

3.3.8.3.2 Hybrid 2. LR-Polypolish

A long-read assembly polished with short reads using Polypolish (Wick & Holt, 2022). Short reads were filtered using fastp (Galaxy v0.19.5) with default settings, which perform quality trimming (Phred < 15), removal of adapter sequences, filtering of reads shorter than 15 bp, and automatic correction of paired-end read overlap and base

errors. Filtered short reads were mapped to long-read assembles using bwa (v0.7.17) installed as part of the Polypolish (v0.6.0) package on an Apple MacBook Pro (Apple M1, OS 14.7). The polypolish_insert_filter.py script was run to filter reads. The filter reads were used to polish the long-read assembly. This process was repeated to complete two rounds of polishing. This fasta file was assessed using checkM (Galaxy v1.0.11).

3.3.8.3.3 Hybrid 3. Uni-Polypolish

A Unicycler (Wick et al., 2017b) assembly polished with short reads using Polypolish. Short reads were filtered using fastp (Galaxy v0.19.5) with default settings, which perform quality trimming (Phred < 15), removal of adapter sequences, filtering of reads shorter than 15 bp, and automatic correction of paired-end read overlap and base errors. Long reads were filtered using Filtlong (Galaxy v0.2.0) with settings Min. length = 1000, and Min. mean quality = 50. Unicycler (Galaxy v0.4.8.0) was used to create an assembly inputting forward and reserve short-read fastqs as well as the long-read fastqs. Filtered short reads were mapped to assemblies using bwa (v0.7.17) installed as part of the Polypolish (v0.6.0) package on an Apple MacBook Pro (Apple M1, OS 14.7). The polypolish_insert_filter.py script was run to filter reads. The filter reads were used to polish the long-read assembly. This fasta file was assessed using checkm (Galaxy v1.0.11).

3.3.8.3.4 Hybrid 4. Uni-Filtered

A Unicycler assembly with long reads filtered using short reads as quality reference. Short reads were filtered using fastp (Galaxy v0.19.5) with default settings, which perform quality trimming (Phred < 15), removal of adapter sequences, filtering of reads shorter than 15 bp, and automatic correction of paired-end read overlap and base errors. Long reads were filtered using Filtlong (Galaxy v0.2.0) with the following settings: minlength = 1000, filtered short reads used as Illumina read reference, Trim non-k-mer-matching activated (removes bases at start and end of sequences not matching *k*-mer), read splitting activated at 500 bases (reads split after 500 consecutive bases fail to match *k*-mer reference). This fasta file was assessed using checkM (Galaxy v1.0.11).

3.3.9 Genome assembly quality control

Through the assembly process checkM (Galaxy v1.0.11) was used to monitor completeness and contamination as polishing steps were applied. At the end of the assembly process, *socru* (Galaxy v2.2.4 (Page et al., 2020)) was used to assess structural integrity, confirming that the chromosome structure had been correctly identified and matched a known orientation. This provided an additional layer of quality control by verifying that the final assembly represented a biologically valid and structurally consistent genome.

3.3.10 Genome annotation

The assemblies were annotated using Prokka (Galaxy v1.14.5 (Seemann, 2014)) and the NCBI Prokaryotic Genome Annotation Pipeline (PGAP), version 2024-07-18.build7555, with default settings and *Salmonella* specified as the genus (Tatusova et al., 2016). A shift from Prokka to PGAP was made during the project, as PGAP provided more conservative but likely more accurate and biologically meaningful annotations. This was evident from a reduction in redundant annotations where Prokka had labelled multiple numbered copies of the same gene.

3.3.11 *In silico* typing and AMR predictions

SeqSero2 (Galaxy v1.2.1 (Zhang et al., 2019)) was used to identify the serovar from genome assemblies and short reads (Zhang et al., 2019). The software program abriTAMR (Galaxy v1.0.14 (Horan et al., 2022)) was used to screen genome assemblies for AMR determinants with the point mutation setting set to *Salmonella*. This software was selected for AMR prediction as it had achieved ISO-certification.

3.3.12 Genome structural analysis

The order and orientation of each sequence file was analysed using *socru* (Galaxy v2.2.4 (Page et al., 2020)) with selected species set to Salmonella_enterica.

3.3.13 Single nucleotide polymorphism analysis

The software snippy4 (Galaxy v4.4.3 (Seemann, 2015)) was used to carry out SNP analysis. Variant calling was carried using hybrid genome assemblies and paired-end short-read data sets. In this chapter, up to 20 isolates were analysed from each of eight distinct stool specimens. For each of the eight stool specimens one isolate was selected to be a within group reference. Each reference was selected based on Illumina sequencing coverage metrics ≥ 58x, ONT sequencing coverage ≥ 25x, a high checkM completeness, a low checkM contamination score, and a solved genome structure via *socru*.

3.3.14 Sequence alignment, read mapping and visualisation

To review SNP calls reads were mapped using minimap2 (Galaxy v2.28). Short reads were mapped to genome FASTA files using preset "sr". Long reads were mapped to genome FASTA files using preset "ava-ont". Mapping was visualised using Artemis (v18.2.0 (Carver et al., 2008)). Gene sequence FASTA files were generated from GenBank files and imported into Artemis. These FASTA files were then used to create sequence alignments with Clustal Omega (www.ebi.ac.uk/jdispatcher).

3.3.15 Hierarchical clustering

Hierarchical clustering was carried out in Enterobase (Zhou et al., 2020). Datasets were filtered by country (United Kingdom) and Lab Contact (Public Health England and Gastrointestinal Bacteria Reference Unit). GrapeTrees were produced using Achtman seven gene MLST and core-genome MLST (cgMLST) V2 + HierCC V1 using the MSTree V2 algorithm.

3.4 Results

3.4.1 Stool specimen and linked metadata

The eight stool specimens used in this project were collected between 13/01/2020 and 08/08/2022. Three stool specimens (20EPA002NSA, 20EPA011NSA, 20EPA012NSA)

were collected before I started this PhD (Table 3.1). These samples were utilised while I was obtaining training and clearance to collect samples from the EPA laboratory. A five-day collection window was implemented for obtaining stool specimens from the EPA. For the duration of the collection window the stool specimen had been stored in a fridge (2-8°C). Prior to submission to the EPA the storage conditions and duration was unknown. Specimens 20EPA002NSA, 20EPA011NSA, 20EPA012NSA, 22EPA051NSA, and 22EPA055NSA were sent from general practitioners (GPs), and specimens 22EPA044NSA, 22EPA053NSA, and 22EPA058NSA were sent from the Norwich and Norfolk University Hospital (NNUH). Seven of the eight stool specimens were submitted by female patients, with the ages of the patients ranging from 2 to 77 years (Table 3.2). Two patients had a recorded travel history: 20EPA002NSA (Thailand) and 22EPA053NSA (South Africa).

3.4.2 Bacterial isolation

Stool specimens were plated onto BSA/XLD plates with a target of ten colonies from each media type. This was achieved with exception of 22EPA055NSA, where colonies only grew on XLD. Additional (20EPA002NSA-21 to -25) colonies were required due to colonies failing screening on MacConkey media (Table 3.3).

Table 3.1: Time Intervals and Bacterial Isolation Data for *Salmonella enterica* Subspecies from Patients

Stool ID	*Salmonella enterica* subsp.	EPA collection date	Quadram collection data	Days between EPA collection and -80°C storage	Bacterial isolation data	Days between storage and bacterial isolation
20EPA002NSA	*Salmonella enterica* subsp. *enterica* serovar Paratyphi B variant Java	13/01/2020	14/01/2020	1	19/04/2022	826
20EPA011NSA	*Salmonella enterica* subsp. *enterica* serovar Paratyphi B variant Java	03/03/2020	03/03/2020	0	19/04/2022	777
20EPA012NSA	*Salmonella enterica* subsp. *enterica* serovar Infantis	21/08/2020	24/08/2020	3	08/06/2022	653
22EPA044NSA	*Salmonella enterica* subsp. *enterica* serovar Typhimurium	18/02/2022	21/02/2022	3	08/06/2022	107
22EPA051NSA	*Salmonella enterica* subsp. *enterica* serovar Enteritidis	20/05/2022	23/05/2022	3	17/10/2022	147
22EPA053NSA	*Salmonella enterica* subsp. *salamae*	25/06/2022	27/06/2022	2	17/10/2022	112
22EPA055NSA	*Salmonella enterica* subsp. *enterica* serovar Anatum	20/07/2022	25/07/2022	5	17/10/2022	84
22EPA058NSA	*Salmonella enterica* subsp. *enterica* serovar Enteritidis	08/08/2022	10/08/2022	2	17/10/2022	68

Table 3.2: Sample Origin and Patient Metadata

Stool ID	Sample origin	Age	Sex	Recent travel	Travel-region
20EPA002NSA	GP	31	Female	Y	Thailand
20EPA011NSA	GP	62	Female	N	n/a
20EPA012NSA	GP	65	Male	N	n/a
22EPA044NSA	Outpatient	2	Female	N	n/a
22EPA051NSA	GP	77	Female	N	n/a
22EPA053NSA	Outpatient	29	Female	Y	South Africa
22EPA055NSA	GP	44	Female	N	n/a
22EPA058NSA	Outpatient	55	Female	no data	no data

GP = General Practitioner

Table 3.3: Isolation of *Salmonella* from Stool Samples by Media Type Used for Colony Selection

Isolate ID	-1	-2	-3	-4	-5	-6	-7	-8	-9	-10	-11	-12	-13	-14	-15	-16	-17	-18	-19	-20
20EPA002NSA	XLD	XLD	XLD	XLD	XLD	XLD	XLD	XLD	XLD	XLD	BSA	BSA	BSA	BSA	BSA	BSA	BSA	BSA	BSA	BSA
20EPA011NSA	BSA	BSA	BSA	BSA	BSA	BSA	BSA	BSA	BSA	BSA	XLD	XLD	XLD	XLD	XLD	XLD	XLD	XLD	XLD	XLD
20EPA012NSA	XLD	XLD	XLD	XLD	XLD	BSA	BSA	BSA	BSA	BSA	XLD	XLD	XLD	XLD	XLD	BSA	BSA	BSA	BSA	BSA
22EPA044NSA	XLD	XLD	XLD	XLD	XLD	XLD	XLD	XLD	XLD	XLD	BSA	BSA	BSA	BSA	BSA	BSA	BSA	BSA	BSA	BSA
22EPA051NSA	BSA	BSA	BSA	BSA	BSA	XLD	XLD	XLD	XLD	XLD	BSA	BSA	BSA	BSA	BSA	XLD	XLD	XLD	XLD	XLD
22EPA053NSA	XLD	XLD	XLD	XLD	XLD	XLD	XLD	XLD	XLD	XLD	BSA	BSA	BSA	BSA	BSA	BSA	BSA	BSA	BSA	BSA
22EPA055NSA	XLD	XLD	XLD	XLD	XLD	XLD	XLD	XLD	XLD	XLD	XLD	XLD	XLD	XLD	XLD	XLD	XLD	XLD	XLD	XLD
22EPA058NSA	XLD	XLD	XLD	XLD	XLD	XLD	XLD	XLD	XLD	XLD	BSA	BSA	BSA	BSA	BSA	BSA	BSA	BSA	BSA	BSA

Isolate ID	-21	-22	-23	-24	-25
20EPA002NSA	XLD	XLD	BSA	BSA	BSA
20EPA011NSA	n/a	n/a	n/a	n/a	n/a
20EPA012NSA	n/a	n/a	n/a	n/a	n/a
22EPA044NSA	n/a	n/a	n/a	n/a	n/a
22EPA051NSA	n/a	n/a	n/a	n/a	n/a
22EPA053NSA	n/a	n/a	n/a	n/a	n/a
22EPA055NSA	n/a	n/a	n/a	n/a	n/a
22EPA058NSA	n/a	n/a	n/a	n/a	n/a

BSA = Brilliance *Salmonella* Agar, XLD = Xylose Lysine Deoxycholate Agar, n/a = not needed

3.4.3 DNA sequencing

Illumina paired-end fastq files and ONT fastq files were run through SeqSero2 to obtain a serovar prediction prior to assembly. Three isolates were removed from the study; 20EPA011NSA_17 failed Illumina sequencing on two occasions, 20EPA012NSA_8 failed ONT sequencing on two occasions, and 22EPA012NSA_19 was removed due to significant ONT read contamination with Typhimurium reads. Sequencing files for 157 isolates were cleared for assembly and further analysis (Table 3.4).

Table 3.4: Number of Isolates with Sequencing Files Cleared for Assembly

Stool ID	Number of isolates
20EPA002NSA	20
20EPA011NSA	19
20EPA012NSA	18
22EPA044NSA	20
22EPA051NSA	20
22EPA053NSA	20
22EPA055NSA	20
22EPA058NSA	20

3.4.4 Issues with Hybrid Assemblies for SNP Analysis

3.4.4.1 Discovering the Issue

Before the main results of this chapter are presented, significant challenges were met while attempting to carry out a SNP analysis using the LR-Pilon (see Section 3.2.8.3.1) assembly pipeline. Observed SNPs ranged from 0-503 in the initial SNP analysis, with ten out of 157 isolates showing more than a 20 SNP distances from the within-group reference. SNPs were observed in *napA* in at least one isolate from all stool specimens. In total, ten different non-synonymous mutations to *napA* were observed across 33 out of 157 isolates; intriguingly these mutations all fell within the same region of the gene. This prompted five putative mutants and five *napA* wildype 22EPA051NSA isolates to be re-sequenced to confirm these mutations. All putative *napA* mutates were not confirmed by Illumina resequencing. This result raised questions of the legitimacy of all SNP calls in the initial analysis.

3.4.4.2 Analysis of LR-Pilon SNP calling

Due to the large number on SNPs in the initial analysis it was not possible to screen all mutations. Mutations observed in the isolates from stool specimen 22EPA051NSA were therefore selected for further assessment. To analyse the problem, gene alignments were made using the LR-Pilon hybrid assembly, an Illumina assembly, and an ONT assembly from the group reference isolate and SNP carrying isolates. This initial screen was to pinpoint the origin of the error, to see if the SNP was present in the Illumina sequence data, the ONT sequence data, or both. Read mapping was used to assess the coverage of regions containing certain SNPs.

3.4.4.2.1 SNP calls in 22EPA051NSA (Enteritidis)

A total of nine SNPs affecting 10 out of 20 isolates were identified for 20EPA051NSA. This included three SNPs in *napA*, two SNPs in *manC1*, and single SNPs in *ydiN*, *dnaJ*, *tldD*, and *rcnA*. DNA sequence alignment for *napA* from the reference isolate 22EPA051NSA_2 and putative *napA* mutants 22EPA051NSA_3, _5, _13 and _14 revealed the Illumina sequencing to be source of SNP. In the LR-Pilon hybrid assemblies, SNPs were located at base positions 204 and 210 in the *napA* gene; there were no SNPs observed at these positions in the *napA* gene from the ONT assembly. The *napA* gene from the Illumina assemblies were littered with SNPs (Figure 3.1). This led me to conclude that the short reads at this location were the cause of the polisher introducing errors into the LR-Pilon hybrid assemblies. Long- and short-read data sets for the reference and putative *napA* mutants were mapped to their corresponding LR-Pilon hybrid assembly. Coverage as low as 2x was observed in the Illumina dataset across the *napA* mutation site in putative mutants again implicating the Illumina data as the source of the SNP.

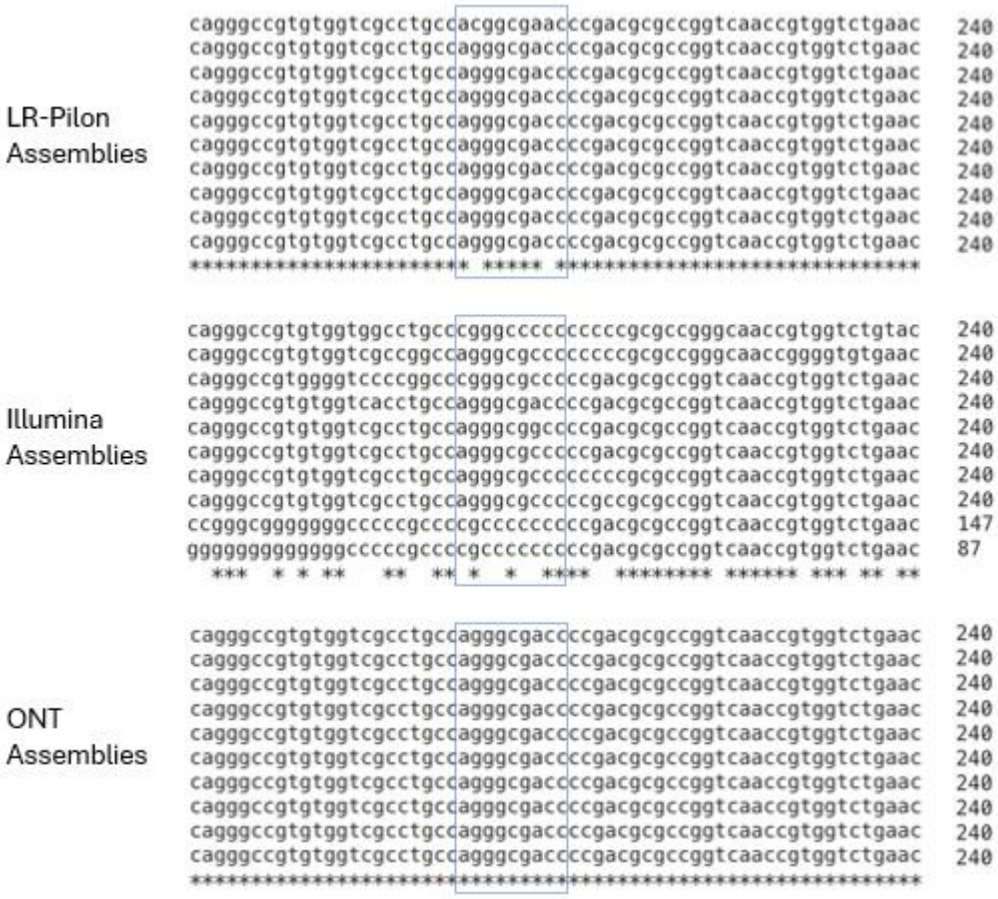

Figure 3.1: Clustal Omega multiple sequence alignments of *napA* from LR-Pilon hybrid assembly, Illumina assembly (short-read only), and ONT assembly (long-read only). Highlighted in blue boxes are the regions within *napA* where putative SNPs were predicted. * = matching base call in all sequences, absent of * = discord between base calls in base position.

Two mutations were observed in the gene *manC1,* both occurring in isolate 22EPA051NSA_15. Base positions 297 and 298 were subject to change from CT-to-GC. In this case the mutation only appears in the 22EPA051NSA_15 hybrid assembly (Figure 3.2).

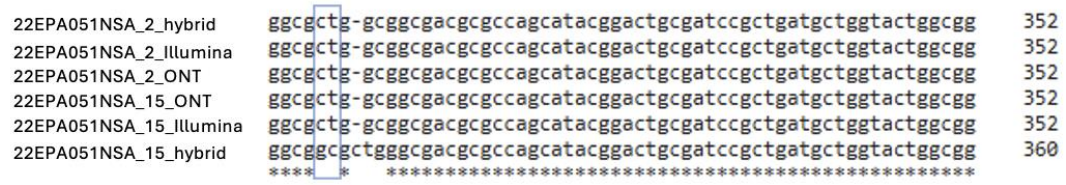

Figure 3.2: Clustal Omega multiple sequence alignments of *manC1* from LR-Pilon hybrid assembly, Illumina assembly (short-read only), and ONT assembly (long-read only). Highlighted in the blue box is the locations within *manC1* where the putative SNP was predicted. * = matching base call in all sequences, absent of * = discord between base calls in base position.

One mutation was called in the *ydiN* gene in isolate 22EPA051NSA_9 a C-to-A substitution at position 1056 in the gene. An alignment of the LR-Pilon hybrid, Illumina, and ONT assemblies revealed a run of G's to be problematic in this region. This run of G's caused a shift in the 22EPA051NSA_9 LR-Pilon hybrid and ONT assembly resulting a frame shift and introduction of a STOP codon. The frame shift did not appear in the 22EPA051NSA_9 Illumina assembly. A STOP codon truncated the reference isolate 22EPA051NSA_2 ONT assembly, this did not appear in the 22EPA051NSA_2 LR-Pilon hybrid or Illumina assembly (Figure 3.3).

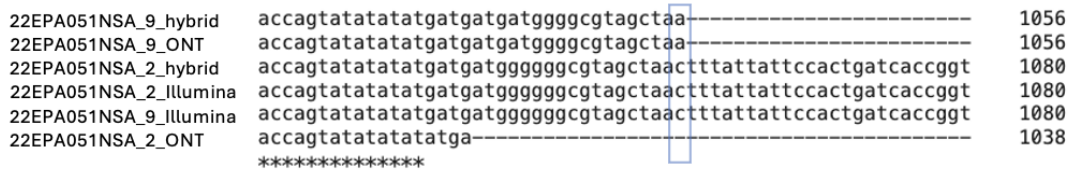

Figure 3.3: Clustal Omega multiple sequence alignments of *ydiN* from LR-Pilon hybrid assembly, Illumina assembly (short-read only), and ONT assembly (long-read only). Highlighted in blue box is the location in *ydiN* where the putative SNP was predicted. * = matching base call in all sequences, absent of * = discord between base calls in base position.

A single SNP was identified in the *dnaJ* gene of two isolates, 22EPA051NSA_6 and 22EPA051NSA_8, with a C-to-T substitution at position 998 (Figure 3.4). This SNP appeared in all assemblies.

Figure 3.4: Clustal Omega multiple sequence alignments of *dnaJ* from LR-Pilon hybrid assembly, Illumina assembly (short-read only), and ONT assembly (long-read only). * = matching base call in all sequences, absent of * = discord between base calls in base position.

A single SNP was identified in *tldD* in isolate 22EPA051NSA_3, a G-to-T substitution at position 961 in the gene (Figure 3.5). This SNP appeared in the three assemblies.

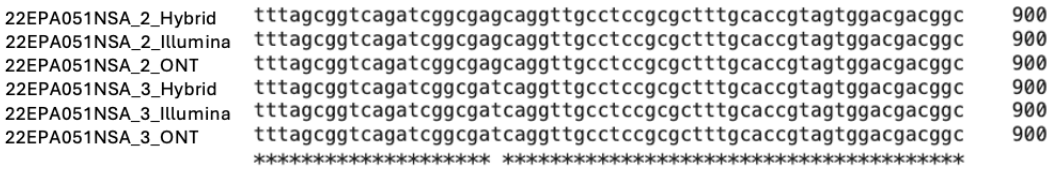

Figure 3.5: Clustal Omega multiple sequence alignments of *tldD* from LR-Pilon hybrid assembly, Illumina assembly (short-read only), and ONT assembly (long-read only). * = matching base call in all sequences, absent of * = discord between base calls in base calls in base position.

Finally, a SNP was observed in *rcnA* in isolate 22EPA051NSA_5, a G-to-A substitution at position 312 (22EPA051NSA_2) in the gene. This SNP appeared in all assemblies (Figure 3.6). This gene was particularly difficult to analyse as the start point of the gene was different for all 22EPA051NSA_5 assemblies.

```
22EPA051NSA_2_Hybrid      ctgagcaccgcgtgctggatgttctggcggacatggcgaggcgagcagcagtggctggcg    354
22EPA051NSA_2_Illumina    ctgagcaccgcgtgctggatgttctggcggacatggcgaggcgagcagcagtggctggcg    354
22EPA051NSA_2_ONT         ctgagcaccgcgtgctggatgttctggcggacatggcgaggcgagcagcagtggctggcg    354
22EPA051NSA_5_Illumina    ctgagcaccgcgtgctgaatgttctggcggacatggcgaggcgagcagcagtggctggcg    356
22EPA051NSA_5_Hybrid      ctgagcaccgcgtgctgaatgttctggcggacatggcgaggcgagcagcagtggctggcg    360
22EPA051NSA_5_ONT         ctgagcaccgcgtgctgaatgttctggcggacatggcgaggcgagcagcagtggctggcg    355
                          ********** ****************************************
```

Figure 3.6: Clustal Omega multiple sequence alignments of *rcnA* from LR-Pilon hybrid assembly, Illumina assembly (short-read only), and ONT assembly (long-read only). * = matching base call in all sequences, absent of * = discord between base calls in base position.

In conclusion, I determined three of the nine SNPs in 22EPA051NSA isolates to be genuine SNPs. SNPs in *napA* appeared to be caused by an issue with the Pilon likely caused by low coverage of Illumina reads in the *napA* gene. The SNPs in *manC1* formed a curious case as the SNPs were not present in the Illumina assembly or ONT assembly, only in the LR-Pilon assembly, suggesting an error in the polishing processing with Pilon. The SNP that appeared in *ydiN* was caused by repetitive sequence that led to the introduction of STOP codons. Here the ONT data appeared to be the source of the error. For *dnaJ*, *tldD*, and *rcnA* SNPs were seen in all three assembly approaches and therefore deemed to be genuine (Table 3.5). Given these results, I do not trust the LR-Pilon pipeline for the purpose of SNP calling between closely matched isolates. This finding highlights the need for careful validation of hybrid polishing pipelines, as miscalls introduced during polishing could lead to incorrect conclusions about genetic variation. In future analyses, SNP identification should rely on high-depth short-read data or validated hybrid approaches specifically optimised for closely related genomes.

Table 3.5: Overview of SNP Calls from 22EPA051NSA LR-Pilon Hybrid Assemblies

Gene	Position in reference	No. isolates affected	Conclusion
manC1	153506	1	Artefact, polishing
manC1	153507	1	Artefact, polishing
ydiN	560097	1	Artefact, repetitive DNA
dnaJ	2333395	2	Real SNP
tldD	3616755	1	Real SNP
rcnA	3960257	1	Real SNP
napA	4680663	1	Artefact, Illumina coverage and polishing
napA	4680667	2	Artefact, Illumina coverage and polishing
napA	4680673	2	Artefact, Illumina coverage and polishing

3.4.4.3 Alternative Hybrid Assembly Approaches

With reasonable suspicion over the accuracy of the LR-Pilon assembly pipeline for precision SNP calling I set out to look at alternative hybrid assembly strategies. A review of the literature revealed some dissatisfaction with the short-read polishing program Pilon (Chen et al., 2021; Wick & Holt, 2022). An alternative program with growing popularity at the time of writing was Polypolish. The substitution of Pilon for Polypolish led to the pipeline LR-Polypolish (See Section 3.2.8.3.2). The LR-Polypolish pipeline was selected to directly test whether changing the short-read polisher would lead to a more accurate SNP analysis. I also decided to test a short-read first assembly approach using Unicycler. Short-read first assemblies were made with two pipelines Uni-Polypolish (See Section 3.2.8.3.3) and Uni-Filtered (See Section 3.2.8.3.4). As a control for this assessment, I used what could be considered the gold standard for SNP analysis which is to use Illumina reads against a reference genome. I selected the Uni-Filtered hybrid assembly to be the reference genome for the short-read SNP calling approach. Finally, I included the ONT long-read only assembly as described in Section 3.2.8.2. This analysis was carried out for all eight sets of isolates (20EPA002NSA, 20EPA011NSA, 20EPA012NSA, 22EPA044NSA, 22EPA051NSA, 22EPA053NSA, 22EPA055NSA, and 22EPA058NSA).

This set of SNP analyses revealed a concerning level of discord between the results when using different hybrid assembly pipelines (Tables 3.6-3.13). Based on the large amount of variability I elected to use short reads for the analysis of SNP diversity presented in Section 3.4.8.

Table 3.6: Estimated Read Coverage of 20EPA002SNSA Isolate Genomes and SNP Calls for Different Assembly Pipelines

ID	Coverage Estimate (x)		Number of SNP calls					
	Illumina	ONT	LR-Pilon	LR-Polypolish	Uni-Polypolish	Uni-Filtered	Illumina	ONT
20EPA002NSA_1	58	43	0	0	10	11	0	43
20EPA002NSA_2	75	61	0	0	4	4	0	53
20EPA002NSA_3	60	90	0	1	5	3	0	56
20EPA002NSA_4	42	95	2	3	33	24	3	62
20EPA002NSA_5	41	80	4	3	21	22	3	41
20EPA002NSA_6	47	29	3	3	28	35	1	105
20EPA002NSA_9	80	24	1	139	1	1	0	5764
20EPA002NSA_12	54	58	3	3	12	12	3	58
20EPA002NSA_13	62	100	3	2	12	11	2	65
20EPA002NSA_15	36	30	6	22	36	28	6	226
20EPA002NSA_16	41	19	503	219	27	30	2	3526
20EPA002NSA_18	54	13	11	32	20	20	1	927
20EPA002NSA_19	63	215	8	1	9	7	0	91
20EPA002NSA_20	46	65	5	8	17	17	3	51
20EPA002NSA_21	76	25	0	0	0	0	0	79
20EPA002NSA_22	76	18	2	13	4	4	2	599
20EPA002NSA_23	68	41	2	3	4	4	3	41
20EPA002NSA_25	65	52	14	4	6	7	3	51
Reference _11	92	64	0	0	0	0	0	0

Table 3.7: Estimated Read Coverage of 20EPA011SNSA Isolate Genomes and SNP Calls for Different Assembly Pipelines

ID	Coverage Estimate (x)		Number of SNP calls					
	Illumina	ONT	LR-Pilon	LR-Polypolish	Uni-Polypolish	Uni-Filtered	Illumina	ONT
20EPA011NSA_1	39	47	19	14	115	30	13	76
20EPA011NSA_2	50	92	6	0	17	15	0	80
20EPA011NSA_3	67	45	16	15	29	30	9	72
20EPA011NSA_4	57	126	17	14	33	32	9	98
20EPA011NSA_5	68	76	18	14	22	21	9	94
20EPA011NSA_6	54	136	18	18	27	24	9	90
20EPA011NSA_7	50	50	20	14	27	27	9	81
20EPA011NSA_8	80	130	17	14	16	16	9	81
20EPA011NSA_9	63	150	17	16	28	27	9	112
20EPA011NSA_10	48	98	19	15	34	32	12	113
20EPA011NSA_12	63	52	17	14	34	32	12	92
20EPA011NSA_13	32	132	29	25	37	40	10	110
20EPA011NSA_14	42	87	2	0	16	17	0	68
20EPA011NSA_15	56	30	19	15	40	40	10	143
20EPA011NSA_16	43	267	5	0	21	20	0	128
20EPA011NSA_18	47	39	3	0	7	7	0	78
20EPA011NSA_19	46	45	26	17	25	25	14	68
20EPA011NSA_20	45	24	3	0	14	14	0	86
Reference _11	70	85	0	0	0	0	0	0

Table 3.8: Estimated Read Coverage of 20EPA012SNSA Isolate Genomes and SNP Calls for Different Assembly Pipelines

ID	Coverage Estimate (x)		Number of SNP calls					
	Illumina	ONT	LR-Pilon	LR-Polypolish	Uni-Polypolish	Uni-Filtered	Illumina	ONT
20EPA012NSA_1	64	87	7	5	6	7	5	755
20EPA012NSA_2	55	73	5	6	5	4	4	719
20EPA012NSA_3	87	90	2	7	3	5	4	329
20EPA012NSA_4	86	59	7	4	5	4	4	689
20EPA012NSA_5	57	81	10	5	4	3	2	723
20EPA012NSA_6	69	60	4	4	5	5	4	717
20EPA012NSA_7	74	24	3	7	2	2	4	484
20EPA012NSA_9	68	30	6	10	5	6	4	570
20EPA012NSA_10	67	82	9	12	8	8	2	976
20EPA012NSA_11	79	43	3	6	2	2	4	467
20EPA012NSA_12	70	62	2	5	3	3	4	370
20EPA012NSA_13	114	19	7	6	5	7	4	1216
20EPA012NSA_15	64	23	2	19	2	2	4	850
20EPA012NSA_16	80	18	6	17	4	3	4	1211
20EPA012NSA_17	65	36	11	5	3	4	4	1159
20EPA012NSA_18	83	32	8	4	4	5	4	758
20EPA012NSA_20	70	56	59	4	7	5	4	317
Reference _14	77	25	0	0	0	0	0	0

Table 3.9: Estimated Read Coverage of 22EPA044SNSA Isolate Genomes and SNP Calls for Different Assembly Pipelines

ID	Coverage Estimate (x)		Number of SNP calls					
	Illumina	ONT	LR-Pilon	LR-Polypolish	Uni-Polypolish	Uni-Filtered	Illumina	ONT
22EPA044NSA_1	75.7	70	2	0	4	4	0	77
22EPA044NSA_2	49.9	50	0	1	23	23	0	73
22EPA044NSA_4	51.1	55	3	0	7	0	0	83
22EPA044NSA_5	51.1	127	0	0	6	7	0	77
22EPA044NSA_6	56.2	112	0	0	4	4	0	59
22EPA044NSA_7	58.6	26	1	1	4	4	0	257
22EPA044NSA_8	80.5	102	0	5	0	0	0	60
22EPA044NSA_9	57.5	74	2	0	4	4	0	64
22EPA044NSA_10	74.2	13	1	164	3	3	0	5269
22EPA044NSA_11	66.2	77	2	0	14	0	0	49
22EPA044NSA_12	73.6	18	2	3	8	9	0	206
22EPA044NSA_14	57.1	18	2	23	0	0	0	1223
22EPA044NSA_15	65.5	35	0	1	1	1	0	93
22EPA044NSA_16	56.5	41	5	1	11	12	0	88
22EPA044NSA_17	69.2	27	2	5	7	8	0	143
22EPA044NSA_18	73.1	106	1	0	10	10	0	74
22EPA044NSA_19	63.9	31	1	1	1	1	0	60
22EPA044NSA_20	62.9	99	1	0	3	3	0	86
Reference _3	65.7	32	0	0	0	0	0	0

Table 3.10: Estimated Read Coverage of 22EPA051SNSA Isolate Genomes and SNP Calls for Different Assembly Pipelines

ID	Coverage Estimate (x)		Number of SNP calls					
	Illumina	ONT	LR-Pilon	LR-Polypolish	Uni-Polypolish	Uni-Filtered	Illumina	ONT
22EPA051NSA_1	74	82	0	0	2	2	0	62
22EPA051NSA_3	56	37	2	7	7	7	1	66
22EPA051NSA_4	59	71	0	0	5	6	0	56
22EPA051NSA_5	66	36	2	1	3	3	1	64
22EPA051NSA_6	60	63	1	1	5	5	1	53
22EPA051NSA_7	77	58	0	0	1	1	0	64
22EPA051NSA_8	76	22	1	1	3	3	1	79
22EPA051NSA_9	76	147	1	0	1	1	0	53
22EPA051NSA_10	59	30	0	0	2	1	0	70
22EPA051NSA_11	49	33	0	0	8	8	0	71
22EPA051NSA_12	59	36	0	0	3	3	0	57
22EPA051NSA_13	77	17	1	0	3	3	0	167
22EPA051NSA_14	64	36	1	0	4	4	0	72
22EPA051NSA_15	70	66	2	0	2	1	0	59
22EPA051NSA_16	70	19	0	0	2	2	0	103
22EPA051NSA_17	78	31	1	0	4	5	0	67
22EPA051NSA_18	72	14	0	1	15	2	0	104
22EPA051NSA_19	73	45	0	0	3	3	0	61
22EPA051NSA_20	63	65	0	0	2	2	0	45
Reference _2	58	164	0	0	0	0	0	0

Table 3.11: Estimated Read Coverage of 22EPA053SNSA Isolate Genomes and SNP Calls for Different Assembly Pipelines

ID	Coverage Estimate (x)		Number of SNP calls					
	Illumina	ONT	LR-Pilon	LR-Polypolish	Uni-Polypolish	Uni-Filtered	Illumina	ONT
22EPA053NSA_1	57	17	1	44	10	10	0	734
22EPA053NSA_2	53	22	2	1	4	4	0	251
22EPA053NSA_3	59	98	0	3	0	1	0	124
22EPA053NSA_4	57	31	1	0	0	0	0	170
22EPA053NSA_5	61	24	2	0	3	3	0	244
22EPA053NSA_6	56	18	1	8	2	2	0	523
22EPA053NSA_7	56	24	0	1	4	4	0	283
22EPA053NSA_8	47	18	3	7	1	1	0	353
22EPA053NSA_9	47	22	0	0	0	0	0	263
22EPA053NSA_10	72	11	2	264	1	1	0	6179
22EPA053NSA_11	54	29	0	2	0	0	0	199
22EPA053NSA_12	62	22	0	1	0	0	0	218
22EPA053NSA_13	73	30	0	0	1	1	0	188
22EPA053NSA_14	62	35	1	0	2	2	0	171
22EPA053NSA_15	70	21	9	56	0	0	0	870
22EPA053NSA_16	69	26	1	0	1	1	0	206
22EPA053NSA_17	84	12	1	59	1	1	0	2024
22EPA053NSA_18	53	16	2	11	1	1	0	733
22EPA053NSA_19	75	79	0	27	0	0	0	289
Reference _20	68	66	0	0	0	0	0	0

Table 3.12: Estimated Read Coverage of 22EPA055SNSA Isolate Genomes and SNP Calls for Different Assembly Pipelines

ID	Coverage Estimate (x)		Number of SNP calls					
	Illumina	ONT	LR-Pilon	LR-Polypolish	Uni-Polypolish	Uni-Filtered	Illumina	ONT
22EPA055NSA_1	62	69	2	4	0	0	1	39
22EPA055NSA_2	77	33	1	1	1	0	0	117
22EPA055NSA_3	66	86	0	1	0	0	0	37
22EPA055NSA_4	74	45	1	1	0	1	0	83
22EPA055NSA_5	52	49	2	4	1	1	1	61
22EPA055NSA_6	62	41	0	1	0	0	0	276
22EPA055NSA_7	79	36	14	19	0	0	0	1226
22EPA055NSA_8	54	24	2	1	0	0	0	49
22EPA055NSA_9	76	82	1	1	0	0	0	38
22EPA055NSA_10	61	51	0	1	8	6	0	85
22EPA055NSA_11	72	69	1	1	6	5	1	50
22EPA055NSA_12	79	47	1	1	4	5	1	46
22EPA055NSA_14	66	21	183	334	94	91	0	646
22EPA055NSA_15	53	40	0	1	6	5	0	55
22EPA055NSA_18	63	55	0	1	2	2	0	79
22EPA055NSA_19	60	33	1	1	1	1	0	302
22EPA055NSA_20	62	40	2	1	1	1	0	104
Reference _13	78	89	1	0	0	0	0	0

Table 3.13: Estimated Read Coverage of 22EPA058SNSA Isolate Genomes and SNP Calls for Different Assembly Pipelines

ID	Coverage Estimate (x)		Number of SNP calls					
	Illumina	ONT	LR-Pilon	LR-Polypolish	Uni-Polypolish	Uni-Filtered	Illumina	ONT
22EPA058NSA_1	55	98	3	0	1	1	1	64
22EPA058NSA_2	69	127	190	0	0	1	1	78
22EPA058NSA_3	59	93	4	0	1	1	1	98
22EPA058NSA_4	54	197	3	3	0	0	0	173
22EPA058NSA_6	74	152	152	0	0	1	1	133
22EPA058NSA_8	67	131	4	0	1	0	1	142
22EPA058NSA_9	68	22	3	0	1	1	1	79
22EPA058NSA_10	53	69	3	1	4	4	1	61
22EPA058NSA_11	49	16	3	1	3	3	1	168
22EPA058NSA_12	60	19	3	1	2	2	1	112
22EPA058NSA_13	60	136	4	0	2	1	1	74
22EPA058NSA_14	59	25	3	0	1	1	1	89
22EPA058NSA_15	66	28	3	0	2	2	1	81
22EPA058NSA_16	48	20	3	0	2	2	1	113
22EPA058NSA_17	55	32	3	0	2	2	1	68
22EPA058NSA_18	58	98	3	1	2	2	1	77
22EPA058NSA_19	54	45	4	0	1	1	1	57
22EPA058NSA_20	64	12	3	8	2	2	2	691
Reference _7	70	79	0	0	0	0	0	0

3.4.5 Genome Level Diversity – Serovar and Sequence Type

A single serovar and sequence type was observed for all isolates recovered from a single patient's stool specimen. The serovar and sequence type per stool specimen is presented in Table 3.14.

Table 3.14: *Salmonella* Classification for Each Stool Specimen

Stool ID	Subspecies	Serovar	Sequence Type
20EPA002NSA	*Salmonella enterica* subsp. *enterica*	Java	ST43
20EPA011NSA	*Salmonella enterica* subsp. *enterica*	Java	ST149
20EPA012NSA	*Salmonella enterica* subsp. *enterica*	Infantis	ST32
22EPA044NSA	*Salmonella enterica* subsp. *enterica*	Typhimurium	ST34
22EPA051NSA	*Salmonella enterica* subsp. *enterica*	Eneritidis	ST11
22EPA053NSA	*Salmonella enterica* subsp. *salamae*	n/a	ST9581
22EPA055NSA	*Salmonella enterica* subsp. *enterica*	Anatum	ST5197
22EPA058NSA	*Salmonella enterica* subsp. *enterica*	Eneritidis	ST11

3.4.6 Genome Level Diversity – Antimicrobial Resistance Determinants

An *in-silico* AMR determinant screen was performed for all isolates. The *mdsA* and *mdsB* genes were identified in all isolates in this study. Additional AMR determinants were identified in *S*. Typhimurium (22EPA044NSA) isolates only. The genes *sul2*, *aph(3")-Ib*, *aph(6)-Id*, and *blaTEM-1* were identified in 19 out 20 *S*. Typhimurium isolates. Isolate 22EPA044NSA_10 did not harbour *sul2*, *aph(3")-Ib*, *aph(6)-Id*, and *blaTEM-1* according to the *in-silico* screen.

3.4.6.1 Lack of AMR Determinants in 22EPA044NSA_10 (Typhimurium)

Alignment of the group reference 22EPA044NSA_3 and 22EPA044NSA_10 revealed the genome location of the AMR genes (*sul2*, *aph(3")-Ib*, *aph(6)-Id*, and *blaTEM-1*) to be within a transposable element. In the reference the AMR determinants were flanked and dissected by IS15DIV transposase insertion sequences. In 22EPA044NSA_10, a single IS15DIV transposase insertion sequence was observed at this site in its genome (Figure 3.7).

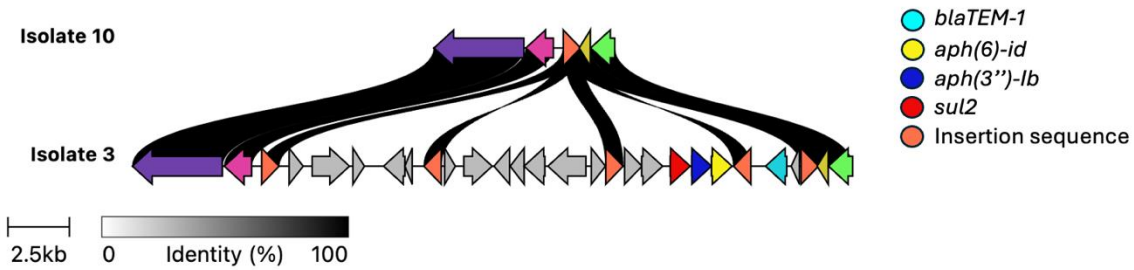

Figure 3.7: Loss of AMR-carrying transposable element in the *S.* Typhimurium genome. Clinker schematic where isolate 3 represents the consensus sequence found in 19/20 isolates from 22EPA044NSA. Flanked by insertion sequences (in orange) several genes including four AMR genes (in red, dark blue, yellow and light blue) were absent in isolate 10. Genes are represented by arrows indicated directionality, with matching colours indicating identical gene sequence. Homology between the two isolates is represented as black bars, regions without black bars linking them are absent in isolate 10.

3.4.7 Genome Level Diversity – Genome Structure Analysis

Structural analysis of the *Salmonella* genomes was carried out using *socru* with hybrid assemblies as input. No variation of genome structure was observed among the isolates from a stool specimen. All isolates observed in the study had Genome Structure (GS) 1.0.

3.4.8 Genome Level Diversity – SNP Analysis

As detailed in Section 3.3.4, I did not consider a SNP analysis from hybrid genomes to be robust due to errors introduced by the assembly pipelines. Here I present the results from a SNP analysis carried out using short reads compared to hybrid genome (Uni-Filtered, See Section 3.2.8.3.4) references.

3.4.8.1 20EPA002NSA – *S.* Java ST 43

SNPs were observed in ten different locations: eight non-synonymous, one synonymous, and one truncation caused by the gain of a STOP codon. Two non-synonymous mutations were observed in gene *dgaR_2*, annotated by Prokka. The NCBI annotation pipeline did not provide an alternative gene name. Both annotation

pipelines identified *dgaR_2* as a transcriptional regulator linked to a sugar phosphotransferase system transporter. The G-to-A mutation was observed in ten isolates, and the T-to-C mutation was observed in three isolates. Non-synonymous mutations were observed in two genes related to biosynthesis of molybdenum cofactor (MoCo). A single mutation in *maoB* in isolate 20EPA002NSA_15, and a single mutation in *maoP* in isolate 20EPA002NSA_20. A STOP codon was introduced into the gene *sicP*, which encodes for Type III secretion chaperone protein sicP. This mutation truncated the protein from 130 amino acids to 46 amino acids in isolate 20EPA002NSA_15. Two isolates (20EPA002NSA_6 and _13) were observed to carry the same non-synonymous mutation in *yciA*. The product of *yciA* is a protein of unknown function predicted to play a role in lipid metabolism or stress response. Additional genes with single mutations included *lysN*, *prfA*, *rseB*, and hypothetical gene *MIEOAJKP_1129* (Table 3.15).

Table 3.15: Summary of Genetic Variants Identified in 20EPA002NSA Isolates

Position in reference	Gene ID	Change identified	Base change	Amino acid change	No. of isolates affected
1007707	*sicP*	Stop gained	G>A	Thr>STOP	1
1165520	*MIEOAJKP_1129*	Non-Synonymous	T>G	Tyr>Asp	10
1259376	*rseB*	Non-Synonymous	A>G	Ser>Gly	1
2151809	*prfA*	Non-Synonymous	T>C	Phe>Leu	1
2194732	*yciA*	Non-Synonymous	C>T	Pro>Ser	2
2337698	*lysN (ydcR)*	Synonymous	C>T	Leu>Leu	1
3105889	*moaB*	Non-Synonymous	G>A	Ser>Asn	1
4018459	*dgaR_2*	Non-Synonymous	T>C	Leu>Pro	3
4018516	*dgaR_2*	Non-Synonymous	G>A	Gly>Asp	10
4690129	*maoP*	Non-Synonymous	A>G	Thr>Ala	1

SNPs in *Gene_1129* and *dgaR_2* divided the 20EPA002NSA isolates into two of clades of ten (Figure 3.8). In total, eight different genomes were separated by at least one core SNP. Three isolates carried the double SNP observed in *dgaR_2*. The isolate 20EPA002NSA_15 was the most genetically distant from the within-group reference, separated by a total of 5 SNPs. Six SNPs was the maximum distance between isolates when using 20EPA002NSA_11 as the reference (Figure 3.8).

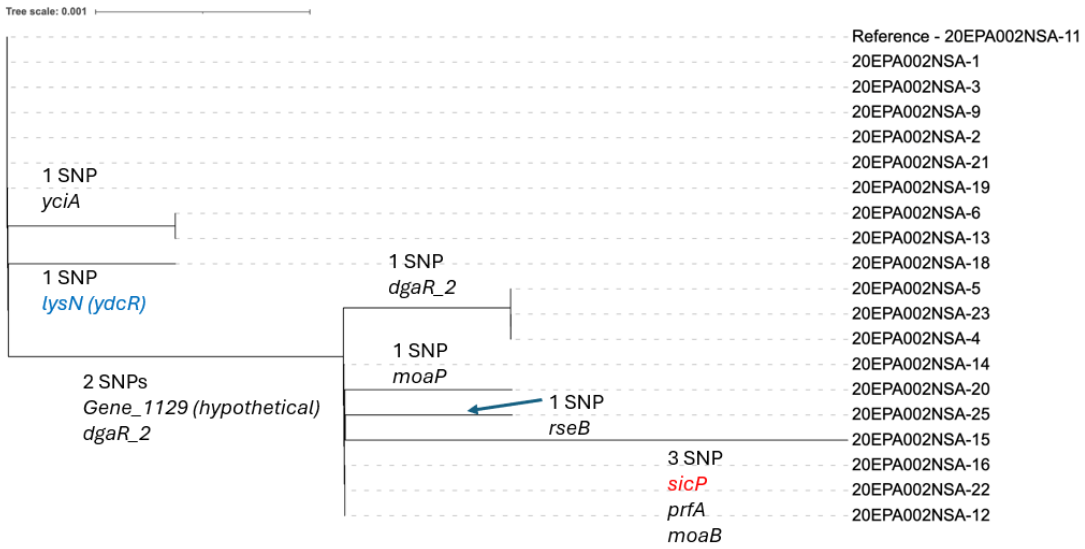

Figure 3.8: Variation between 20EPA002NSA Isolates. Core genome maximum likelihood tree for twenty *S.* Java ST 43 isolates. Tree overlaid with the SNPs responsible for each branch. Key for SNP type: black = non-synonymous, blue = synonymous, and red = STOP gained.

3.4.8.2 20EPA011NSA – *S.* Java ST 149

Core SNPs were observed in nineteen different locations, twelve non-synonymous, five synonymous, and two in non-coding DNA (Table 3.16). A set of six SNPs separate a group of 13 isolates from the reference (Figure 3.9). This set of six SNPs include four non-synonymous and two synonymous SNPs. Non-synonymous SNPs in this grouping were found in *secY*, a core component of the Sec translocon; *dmsC_3*, a subunit of the dimethyl sulfoxide (DMSO) reductase complex; *sfmF*, a fimbrial subunit involved in fimbriae assembly; and *cnoX*, a redox-active protein crucial for stress response. The group of thirteen isolates further divides into three more groups (Figure 3.9). A group of eleven isolates share three SNPs, two non-synonymous and one synonymous, 20EPA011NSA_13 has three unique SNPs, and 20EPA011NSA_19 has seven unique SNPs.

Table 3.16: Summary of Genetic Variants Identified in 20EPA011NSA Isolates

Position in reference	Gene ID	Change identified	Base change	Amino acid change	No. of isolates affected
105760	*recG*	Synonymous	G>A	Leu>Leu	1
224107	*pucK*	Synonymous	T>C	Gly>Gly	1
238616	*bcsA*	Non-Synonymous	C>T	Ala>Val	1
364934	Non-coding	-	-	-	1
389465	*OLPEJMNH_343*	Non-Synonymous	A>G	Gln>Arg	1
448012	*secY*	Non-Synonymous	G>A	Arg>Gln	13
498868	*tldD*	Non-Synonymous	A>G	Gln>Arg	1
619790	*tsar* (*tdcA*)	Non-Synonymous	T>C	Val>Ala	1
1658795	*setB*	Synonymous	A>G	Leu>Leu	11
2375311	*OLPEJMNH_2267*	Non-Synonymous	G>A	Val1>Ile	11
2390649	*OLPEJMNH_2281*	Synonymous	T>C	Ser>Ser	13
2716374	*ptsG*	Non-Synonymous	A>G	Ser>Gly	1
2907880	*dmsC_3*	Non-Synonymous	C>T	Ala>Val	13
2961186	*ydcV_2* (*potI*)	Synonymous	C>T	Leu>Leu	13
3302661	*sfmF*	Non-Synonymous	G>A	Gly>Asp	13
3348459	*cnoX*	Non-Synonymous	G>A	Ala>Thr	13
3831410	Non-coding	-	-	-	1
4381614	*rpoB*	Non-Synonymous	C>T	Pro>Leu	11
4448804	*hslU*	Non-Synonymous	G>A	Gly>Ser	1

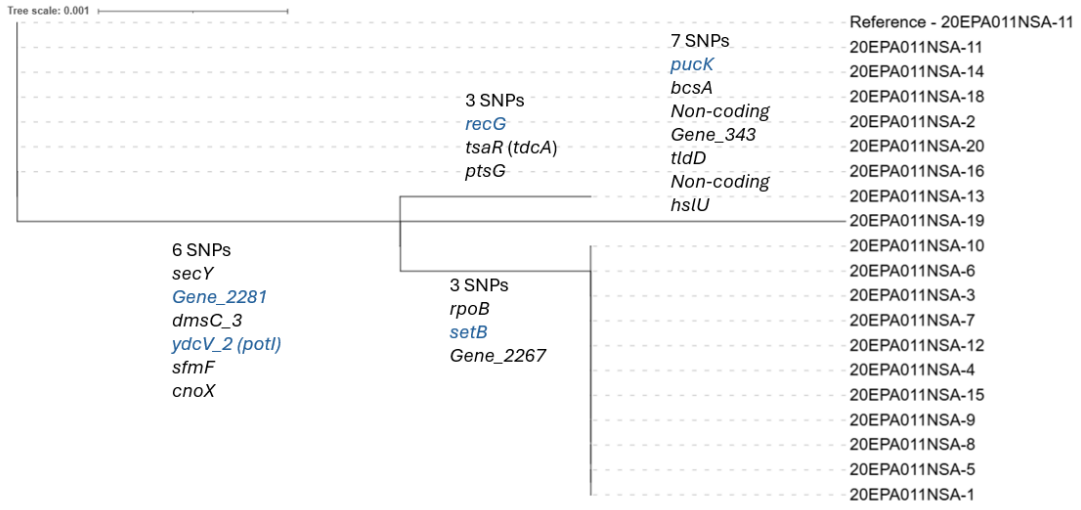

Figure 3.9: Variation between 20EPA011NSA Isolates. Core genome maximum likelihood tree for nineteen *S.* Java ST 149 isolates. Tree overlaid with the SNPs responsible for each branch. Key for SNP type: black = non-synonymous and blue = synonymous.

3.4.8.3 20EPA012NSA – *S*. Infantis ST 32

Three SNPs were observed affecting four isolates. Isolates 20EPA012NSA_4 and _10 had a non-synonymous mutation in *bigA*. The *bigA* gene encodes a hypothetical surface protein. Single non-synonymous mutations were observed in the *hilA* gene of 20EPA012NSA_14, and the *ompN_1* gene of 20EPA012NSA_1 (Table 3.17). The *hilA* gene encodes a transcriptional regulator which acts as the master regulator of *Salmonella* pathogenicity island 1 (SPI-1). The *ompN* genes encodes an outer membrane protein associated with passive transport of small molecules across the outer membrane.

Table 3.17: Summary of Genetic Variants Identified in 20EPA012NSA Isolates

Position in reference	Gene ID	Change identified	Base change	Amino acid change	No. of isolates affected
407020	*bigA*	Non-Synonymous	A>T	Glu>Asp	2
1039275	*hilA*	Non-Synonymous	A>G	Asn>Ser	1
1906318	*ompN_1*	Non-Synonymous	C>T	Leu>Phe	1

3.4.8.4 22EPA044NSA – *S*. Typhimurium ST 34

No core SNPs were observed in the 22EPA044NSA isolates.

3.4.8.5 22EPA051NSA – *S*. Enteritidis ST 11

Three SNPs affecting four isolates were observed. Isolates 22EPA051NSA_6 and _8 had a non-synonymous mutation in *dnaJ*. A STOP codon was introduced into the *rcnA* gene in 22EPA051NSA_5, truncating the gene from 284 amino acids down to 107 amino acids. The *rcnA* gene encodes a membrane-bound efflux protein involved in exporting excess nickel and cobalt ions, supporting metal ion homeostasis. A single mutation was observed in *tldD* in 20EPA051NSA_3 (Table 3.18).

Table 3.18: Summary of Genetic Variants Identified in 22EPA051NSA Isolates

Position in reference	Gene ID	Change identified	Base change	Amino acid change	No. of isolates affected
502845	*tldD*	Non-Synonymous	G>T	Glu>Asp	1
846347	*rcnA*	Stop gained	G>A	Trp>STOP	1
3905116	*dnaJ*	Non-Synonymous	C>T	Pro>Leu	2

3.4.8.6 22EPA053NSA – *S. salamae* ST 9581

No core SNPs were observed in the 22EPA053NSA isolates.

3.4.8.7 22EPA055NSA – *S.* Anatum ST 5197

Two core SNPs were observed in two individual isolates, a gain of a STOP codon in 22EPA055NSA_5 and a synonymous mutation in 22EPA055NSA_17 (Table 3.19). The STOP gained in *rcsC* truncates the gene from 948 amino acids to 237 amino acids. The *rcsC* gene encodes a sensor histidine kinase that is a key component of the RcsCDB phosphorelay system. This system plays a crucial role in regulating the synthesis of capsular polysaccharide and modulating motility.

Table 3.19: Summary of Genetic Variants Identified in 22EPA055NSA Isolates

Position in reference	Gene ID	Change identified	Base change	Amino acid change	No. of isolates affected
1543015	*rcsC*	Stop gained	C>T	Gln>STOP	1
2080250	*adhE_4*	Synonymous	G>T	Val>Val	1

3.4.8.8 22EPA058NSA – *S*. Enteritidis ST11

A single non-synonymous SNP was observed in a single isolate, 22EPA058NSA_20 (Table 3.20). The *ftsK* gene encodes a protein that is essential for cell division and chromosome segregation.

Table 3.20: Summary of Genetic Variants Identified in 22EPA058NSA Isolates

Position in reference	Gene ID	Change identified	Base change	Amino acid change	No. of isolates affected
2965035	*ftsK*	Non-synonymous	G>T	Val>Leu	1

3.4.9 Hierarchical clustering of *Salmonella* Java

To explore the significance of the SNP distances observed among *S.* Java isolates within 20EPA002NSA and 22EPA011NSA, hierarchical clustering was performed within Enterobase. *S*. Java infection is common in the UK frequently appearing in the top 10 serovars causing foodborne illness each year (Figure 3.10). The most frequently observed sequence type is ST43. Observation of illness caused by *S. Java* in the UK was notably on the rise from 2012 through 2019. The COVID-19 pandemic disrupted this pattern during 2020-2021. From 2022-2024 the rise in cases was observed once again. Intriguingly, the pandemic may have resulted in the loss of ST42 from circulation in the UK.

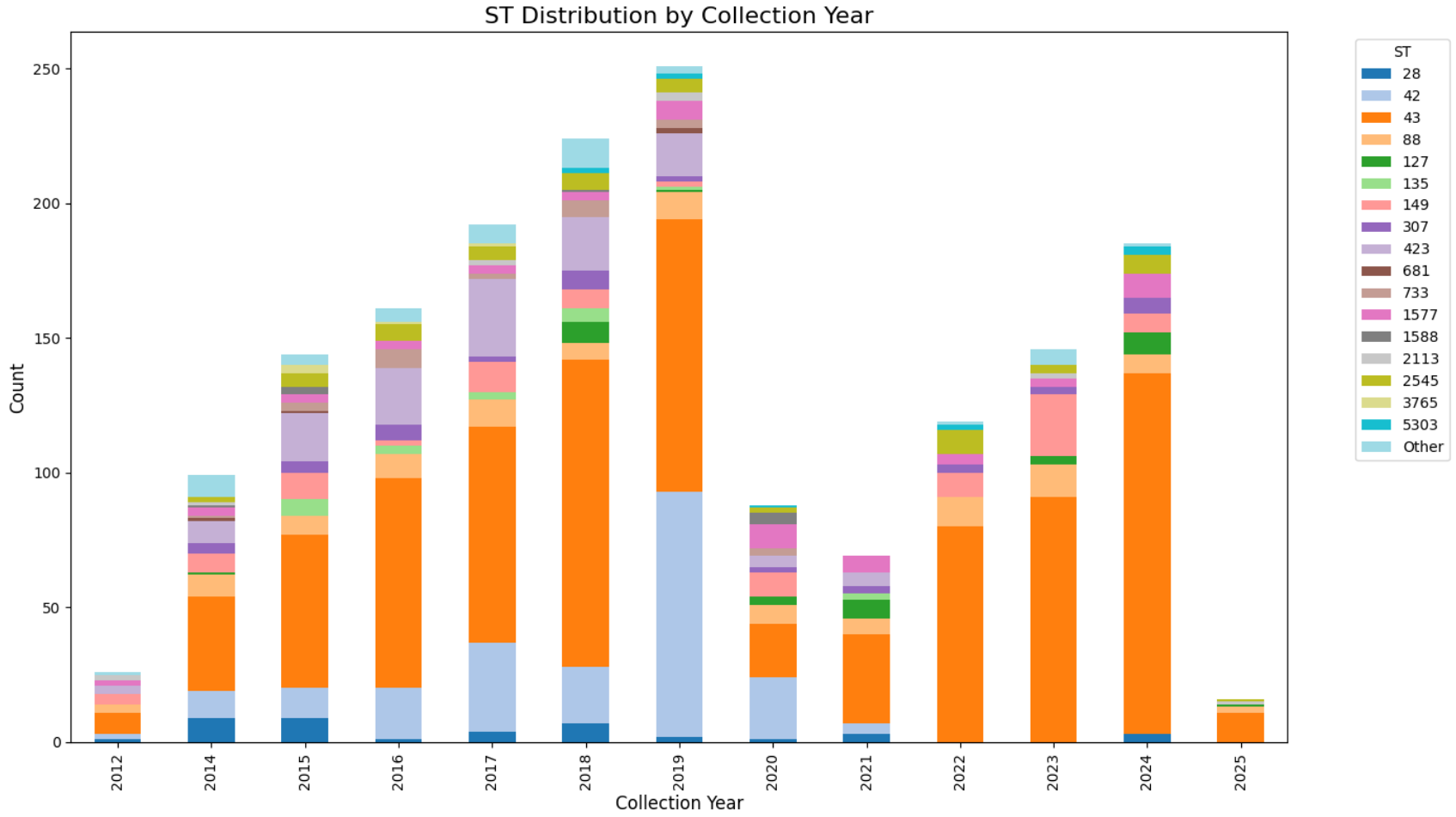

Figure 3.10: Stacked bar chart displaying the count of Serotype (ST) by collection year

The *S.* Java population observed in the UK is dominated by ST43, yet exhibits considerable diversity, with the presence of several other sequence types, notably ST42, ST423, ST88, ST149, ST1577, and ST2545. The two STs linked to the isolates in this project were ST43 (20EPA002NSA) and ST149 (20EPA011NSA). As previously mentioned, ST43 forms the largest cluster of isolates observed in the UK, whilst ST149 ranks fifth in the order of most frequent during the time period January 2012 to February 2025. ST149 forms the central hub of a 7 gene MLST grapetree which is suggestive that this ST149 presents a key or ancestral ST in the population (Figure 3.11). However, this does not hold true when visualising a higher resolution cgMLST grapetree where ST42 forms the central hub, with ST43 forming a diverse cluster from which ST149 is a branch.

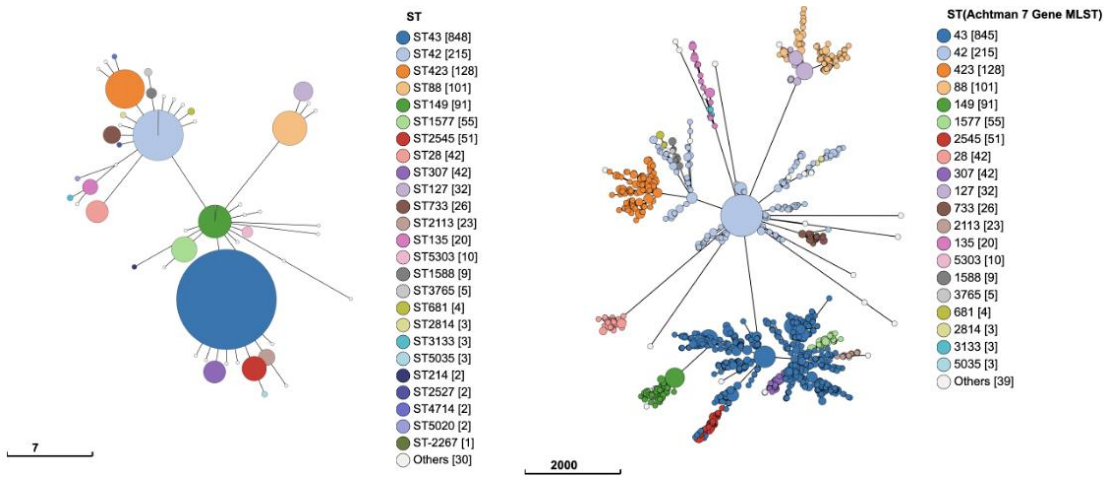

Figure 3.11: GrapeTrees showing UKHSA United Kingdom origin *S.* Java isolates. Left: An Achtman 7 Gene MLST GrapeTree with the key displaying ST based on the Achtman 7 Gene MLST scheme. Right: A cgMLST GrapeTree with the key displaying ST based on the Achtman 7 Gene MLST scheme. ST43 is shown in dark blue, and ST149 in dark green. The scale bar represents a distance in alleles.

To evaluate the clustering patterns, isolates were screened at hierarchical clustering (HC) thresholds of HC5 and HC10. HC5 and HC10 refer to HC cluster levels in EnteroBase that group isolates differing by no more than 5 or 10 cgMLST alleles, respectively. In many *Salmonella* datasets, these allele-difference thresholds correspond approximately to maximum pairwise distances of around 5 SNPs (HC5) or 10 SNPs (HC10), although the exact SNP equivalents vary by dataset and analysis pipeline (Mook et al., 2018; Zhou et al., 2020). These thresholds define the size of

clusters relevant for tracing transmission pathways and identifying closely related isolates within a population. Clusters identified at HC5 and HC10 are considered epidemiologically relevant because they represent groups of isolates that are genetically close enough to potentially share similar virulence characteristics, resistance profiles, and transmission routes (Chattaway et al., 2019a; Zhou et al., 2020). To perform hierarchical clustering for 20EPA002NSA and 20EPA011NSA, the Illumina paired-end files for both the group reference isolate and selected isolates with a high SNP distance from the reference were uploaded to Enterobase. For 20EPA002NSA this was 20EPA002NSA_11 (reference), 20EAP002NSA_4 (3 SNP distance) and 20EPA002NSA_15 (5 SNP distance). For 20EPA011NSA this was 20EPA011NSA_11 (reference), 20EPA011NSA_1 (9 SNP distance), 20EPA001NSA_13 (9 SNP distance) and 20EPA011NSA_19 (13 SNP distance). A difference was observed for 20EPA002NSA at HC5: 20EPA002NSA_11 and 20EPA002NSA_4 had a HC5 of 224147 while 20EPA002NSA_15 had a HC5 of 520667. The three 20EPA002NSA isolates clustered together at HC10 (7172).

A difference was observed between the 20EPA011NSA isolates at HC5 and HC10, with all isolates clustering together at HC20 of 21039. Specifically, for 20EPA011NSA_11, the HC5 and HC10 values were 520186, while for 20EPA011NSA_1, the HC5 and HC10 values were 520043 (Figure 3.12). Isolates 20EPA011NSA_13 and 20EPA011NSA_19 shared the same HC10 cluster (20139) but differed at HC5 with values of 526068 and 5206066, respectively. This result suggests that the separation seen in the 20EPA011NSA isolates (Figure 3.9) is large enough to be deemed epidemiological relevant and that for *S*. Java ST149 a single colony is not sufficient to capture the genome level diversity present.

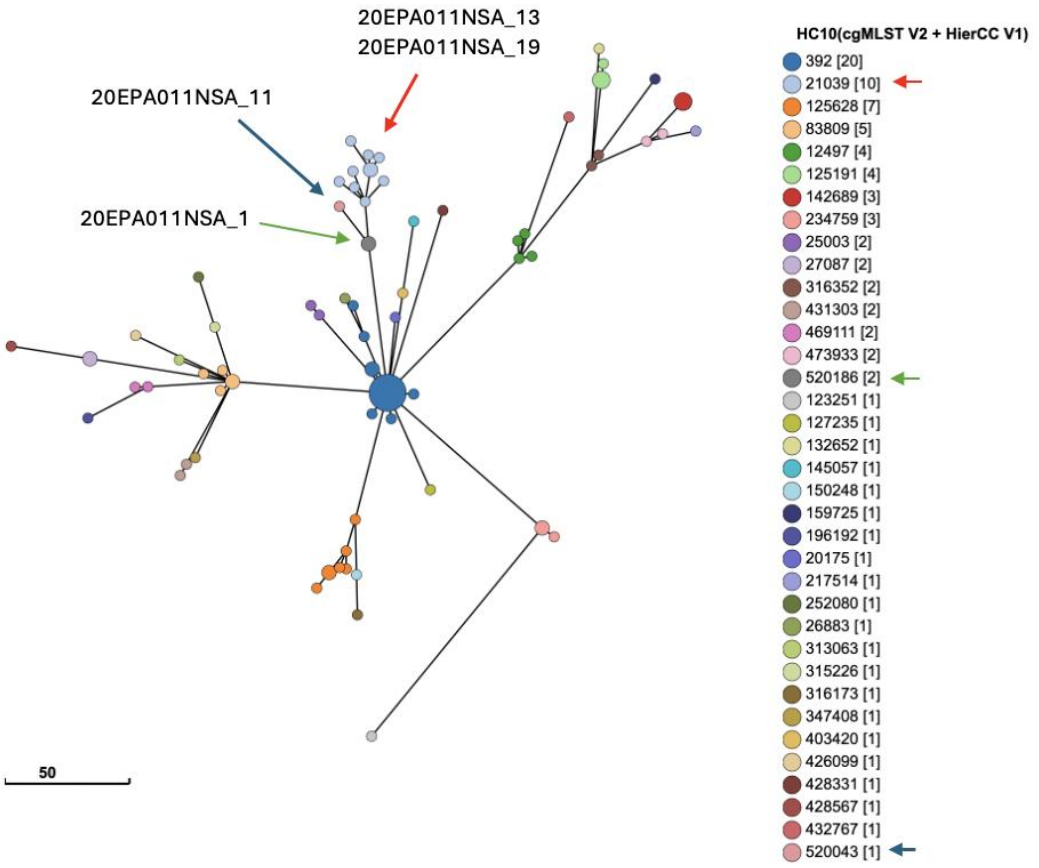

Figure 3.12: GrapeTree showing cgMLST for UKHSA United Kingdom origin *S.* Java ST149 isolates supplemented with 20EPA011NSA_1 and 20EPA011NSA_11. The key is at the HC10 levels. The 20EPA011NSA_1 containing cluster is marked with a green arrow, the 20EPA011NSA_11 containing cluster is marked with a blue arrow, and the 20EPA0011NSA_13 and _19 cluster is marked with a red arrow. The scale bar represents a distance in alleles.

3.5 Discussion

This chapter utilised hybrid genome sequencing to explore genome-level diversity among isolates recovered from a single patient's stool specimen, who presented with gastroenteritis symptoms in Norfolk, UK. The findings were analysed within an epidemiological framework to evaluate the validity of using a single bacterial colony for accurately capturing relevant information for outbreak detection. A key consideration when interpreting the findings of this study is the sample size. Only eight patient stool specimens were examined, each providing up to twenty *Salmonella* isolates for detailed analysis. While this design enabled an in-depth exploration of within-host genomic diversity, the limited number of patients constrains the ability to generalise the findings to broader populations. The study was intended as an exploratory investigation to assess the feasibility and value of multi-isolate sequencing from single

infections. Nevertheless, the insights gained provide important preliminary evidence that can guide future large-scale studies. Increasing the number of patient samples, sampling over time, and incorporating different clinical and epidemiological contexts would allow stronger statistical inference and help determine whether the genomic patterns observed here are representative of wider Salmonella diversity dynamics.

3.5.1 Hybrid assembly

My perspective on hybrid genome assembly is that it serves as a means to an end rather than an ideal solution. The challenges I faced during this chapter have led me to strongly believe that sequence data from different platforms should be analysed independently whenever possible, as the most reliable data is often that which has been least processed by algorithms. From the dataset of long and short reads generated for this chapter, it became apparent that a thousand different results could be produced by combining various assemblers and polishers in different orders, each with parameters that can be tweaked. This variability is concerning as a researcher aiming to screen for diversity between isolates that may or may not be identical.

This study highlights the ongoing need for improvements in sequencing technology, even for high-resolution analyses of small bacterial genomes. A key limitation identified in this work was the R9 ONT chemistry. Assemblies generated exclusively with ONT data were highly error-prone, sometimes exhibiting thousands of SNPs compared to their group reference. In contrast, assemblies based solely on Illumina reads revealed no SNPs for the same samples. Worryingly, the high error rate in ONT reads was not always the sole issue; dropout in Illumina read coverage also contributed to errors in the hybrid assemblies. These findings highlight the constraints of existing sequencing technologies and software packages, emphasising the necessity for enhanced methods to ensure precise genomic analyses.

Despite the variability associated with hybrid genome assemblies, their importance in identifying structural elements of bacterial genomes cannot be understated. The combined advantages of short reads, which offer high base-level accuracy, and long reads, which cover repeating regions and structural variations, are valuable (Luan et al., 2024; Wick et al., 2017a). This synergy is particularly valuable for resolving complex

genomic features such as prophages, insertion sequences, and large chromosomal rearrangements (Elek et al., 2023; Huisman et al., 2022; Waters et al., 2025), which are often critical to understanding bacterial evolution, pathogenicity, and antimicrobial resistance. Structural regions like this are inadequately addressed using short-read or long-read assemblies when used in isolation. By combining the datasets, hybrid assemblies can provide a more comprehensive representation of the genome. This makes hybrid genomes an essential tool for high-resolution analyses of bacterial genomes where structural insights are pivotal(Wick & Holt, 2022). These capabilities are especially important in epidemiological scenarios, where understanding an isolate's genomic architecture might indicate its prospective virulence and resistance profile.

ONT's R10 chemistry and PacBio's HiFi sequencing offer promising solutions to overcome the limitations of hybrid assemblies by providing long reads with improving base-level accuracy. R10 chemistry improves the ability to resolve repetitive regions, while PacBio HiFi reads combine the long-read capability with an accuracy comparable to short reads (Bogaerts et al., 2024; Zidane et al., 2025). These advancements reduce the dependency on hybrid approaches. Enabling researchers to generate high-quality assemblies that capture both genomic structure and sequence accuracy in a single dataset. Such technologies are poised to revolutionise bacterial genome analysis by streamlining workflows and minimising potential sources of error. An example of the potential of ONT R10 is presented in the direct sequencing of *Campylobacter* from stool in Chapter 5.

3.5.2 Genome Level Diversity

Understanding the genetic diversity within a single infection is crucial for accurately characterising pathogen behaviour, virulence, resistance mechanisms, and outbreak dynamics. Mixed infections, where multiple strains or sequence types of the same species coexist, can significantly influence treatment outcomes and epidemiological tracing (Balmer & Tanner, 2011; Liu et al., 2015). However, the common practice of analysing single colonies from a clinical sample risks overlooking this diversity. This could lead to incomplete or biased conclusions about the infection becoming particularly problematic for outbreak detection, where advanced methods such as hierarchical clustering and SNP analysis are used to identify related strains

(Chattaway et al., 2023). By focusing on a single colony the ability to detect subtle genomic variations that link strains to a common source or differentiate unrelated cases may be compromised. Accounting for the entire genetic landscape of an infection improves the resolution of hierarchical clustering and SNP analysis, allowing for more precise outbreak investigations and public health responses.

3.5.2.1 Typing

No mixed infections were observed at the species, serovar, or sequence type level. This finding is supported by the rarity in which mixed *Salmonella* serovar or sequence type infections have been reported in the literature (Authority, 2021; Mank et al., 2010).

3.5.2.2 AMR

Variations in AMR profiles among isolates from the same patient were uncommon, for seven of the eight patient stool samples, the *Salmonella* isolates exhibited minimal AMR determinants, carrying only the efflux pump genes *mdsA* and *mdsB* (Song et al., 2015). However, a striking exception was observed in the *S.* Typhimurium infection from 22EPA044NSA. While one isolate was predicted to be sensitive, the consensus AMR profile included *sul2*, *aph(3'')-Ib*, *aph(6)-Id*, and *blaTEM-1*. Long-read sequencing facilitated the assembly of full circular genomes, allowing for a thorough analysis of the genome sequences, including the precise positions of AMR genes within repetitive regions. This analysis revealed that the four AMR genes were clustered within a genomic region flanked by five IS15DIV insertion sequences. In the sensitive isolate a single IS15DIV insertion sequence was identified at this location, none of the four mentioned AMR genes, highlighting the role of mobile genetic elements in the dissemination and loss of resistance determinants in this set of isolates. Detecting varying *Salmonella* AMR profiles within a single infection highlights that analysing a single colony may not accurately reflect the broader *Salmonella* population responsible for the infection, this can compromise investigation conclusions. If a sensitive colony were chosen as representative the presences of *sul2*, *aph(3'')-Ib*, *aph(6)-Id*, and *blaTEM-1* would have been missed. While sequencing multiple isolates individually enhances the depth of investigation, the number of isolates required to fully capture the genetic diversity of a *Salmonella* population during an infection remains unclear. For *Campylobacter* it has been suggested that up to 80 isolates

would be needed to capture 95% of core non-recombinant SNPs (Djeghout et al., 2022).

3.5.2.3 SNP and Hierarchical clustering

The SNP analysis identified SNPs in isolates from six of the eight stool specimens screened in this chapter. In contrast, the isolates from 22EPA044NSA (*S*. Typhimurium) and 22EPA053NSA (*S*. salamae) did not contain any SNPs and can be considered clonal. While isolates from four stool specimens carried SNPs, no single isolate contained more than one SNP. From a SNP analysis and clustering perspective, these single-SNP isolates do not significantly impact downstream epidemiological analyses, suggesting that analysing more than one colony may not be necessary. However, concerns arise with the *S*. Java isolates, particularly those from 20EPA011NSA. The presence of four unique HC5 clusters and three HC10 clusters among these isolates underlines possible difficulties in conducting epidemiological analysis using a single colony.

Sequencing multiple isolates per patient sample incurs significant costs, including labour, colony isolation, DNA extraction, sequencing, and data storage. These factors make this approach impractical for large-scale diagnostic pipelines. Alternative strategies have been developed in an attempt to capture population-level information in a cost friendly manner, namely sweep sequencing and pool-seq (Holt et al., 2009; Mäklin et al., 2021). These methods sequence multiple isolates by combining DNA extractions and sequencing them as a single sample. While these methods address the limitations of analysing a single colony, they have their own challenges. Sweep sequencing can obscure minor alleles, resulting in a skewed representation of genetic diversity. Pool-seq is captures a broader population diversity but can introduce biases due to unequal DNA contributions from individual isolates, and low-frequency variants may remain undetected without sufficient sequencing depth.
Metagenomics is a direct sequencing approach enabling sequencing without the need for isolating colonies. This method offers a comprehensive view of the microbial population, capturing genetic diversity at the population level and has the potential to detect minor alleles (Olm et al., 2021; Vicedomini et al., 2021). Currently metagenomics is the most expensive solution and is hampered by other

computational challenges including resolving individual genomes and linking specific AMR and plasmid profiles or SNPs to individual strains. Additionally, low-abundance species or variants may be underrepresented if the sequencing depth is insufficient.

3.5.2.4 Recurring SNPs

3.5.2.4.1 20EPA002NSA – *S*. Java ST43

Some SNPs were observed in more than one isolate from within a single specimen. The most interesting being a double SNP within the *dgaR_2* gene with one of the SNPs occurring in ten *S*. Java ST43 (20EPA002NSA) isolates and the double SNP occurring in three of these isolates. The *dgaR_2* gene is predicted to be a transcriptional regulator of a sugar PTS transporter. Both SNPs result in missense variants within the PTS EIIA mannose/sorbose-specific type-4 domain of the protein. Mutations in the EIIA domain could affect the regulation of sugar uptake. The EIIA protein regulates the activity of the PTS by interacting with other components in the system and can be involved in signal transduction pathways (Miller et al., 2013). A mutation could cause an unregulated or inefficient transporter, which could be detrimental or beneficial depending on the specific environment. In some cases, mutations could make the transporter hyperactive, enabling the bacterium to take up sugars more efficiently, which could be an adaptive advantage in nutrient-limited environments (Warsi et al., 2018). A mutation that occurs in combination with *dgaR_2* is in *Gene_1129*, a hypothetical gene found with an operon of phage related genes.

A distinct mutation identified in two 20EPA002NSA isolates, independent of the dgaR_2 mutation, involves a single SNP in *yciA*. This gene is predicted to encode an intracellular septation protein associated with lipid metabolism. Previous studies have shown *yciA* to be induced during pig infections (Huang et al., 2007) and repressed under sodium hypochlorite stress (Li et al., 2022), underscoring its sensitivity to environmental cues. These findings suggest that *yciA* may play a key regulatory or functional role in environmental adaptation. Collectively, the mutations observed across the 20EPA002NSA isolates may reflect distinct adaptive strategies under varying selective pressures.

3.5.2.4.2 20EPA011NSA – *S*. Java ST149

The 20EPA011NSA isolates form an interesting case as there is a clear split in the population screened. Six shared SNPs separate thirteen isolates form the remaining seven, and three shared SNPs further separate eleven within the group of thirteen. The group of six SNPs is comprised of four non-synonymous and two synonymous changes. The non-synonymous SNPs are in *secY*, *dmsC_3*, *sfmF*, and *cnoX*. These genes have predicted functions that appear valuable in niche adaption. The *dmsC_3* is likely to be a paralog of the *dmsC* gene, a DMSO reductase involved in anaerobic respiration, which could enhance *Salmonella*'s survival in anaerobic environments, such as the intestinal lumen in a gain-of-function scenario (GoF) (Cruz et al., 2023). Conversely for this gene loss-of-function (LoF) might reallocate energy away from unused systems - a strategy seen in human-adapted strains (McClelland et al., 2004). Orthologs of *dmsA* (a partner of DmsC) are intact in non-typhoidal strains but are pseudogenes in *S*. Typhi and Paratyphi, and the DmsABC pathway has accrued inactivating mutations in human-adapted typhoidal serovars (J. S. Kim et al., 2024). This suggests that losing DMSO reductase activity can be tolerated or even advantageous in certain niches. The *cnoX* gene encodes a thiol-dependent peroxidase and chaperone that protects bacteria from oxidative stress which plays a protective role during host infection by countering oxidative bursts from immune cells, thereby enhancing *Salmonella*'s ability to evade immune defences (Dupuy & Collet, 2021). LoF appears unlikely for *cnoX* as during epithelial invasion and within intracellular vacuoles reduced protection to oxidative stress would have a negative impact on survival unless redundant systems are compensating or CnoX is not expressed in specific niches. GoF in CnoX would be advantageous for *Salmonella* enhancing survival under oxidative stress, especially in macrophages. The product of the *sfmF* gene is thought to be a chaperone or accessory protein that assists in fimbrial biogenesis which may support attachment to host epithelial cells, enhancing virulence and biofilm formation (Guo et al., 2009; Meysman et al., 2013). GoF in SfmF could benefit colonisation by promoting adhesion and persistence offering advantages in the intestinal lumen and during epithelial invasion. LoF may aid immune evasion or reflect niche specialisation (e.g., Typhi losing fimbriae) (McClelland et al., 2004; Yue et al., 2012). Finally, the *secY* gene encodes a core component of the Sec translocon, a protein-conducting channel in the bacterial membrane. It is essential for bacterial viability as it facilitates the secretion of virulence factors critical for host invasion and membrane protein assembly critical for

survival (Durack et al., 2015; Oswald et al., 2021). GoF for SecY may enhance secretion of virulence proteins in invasive and intracellular contexts. A point mutation in SecY can act like a prl (protein localisation) suppressor, expanding the range of secreted proteins. In *Listeria*, a single SecY mutation restored secretion of key effectors and increased virulence by broadening SecY's substrate specificity (Durack et al., 2015). LoF for SecY would likely be deleterious across all human niches (Oswald et al., 2021).

Taken together, these observations hint at potential functional divergence that may reflect adaptive processes acting within the host or during transmission. However, given the small number of isolates analysed, these patterns should be interpreted cautiously. The apparent clustering could arise from stochastic variation or limited sampling rather than genuine selective pressure. Further comparative analyses across a broader isolate set would be required to determine whether these mutations represent true adaptive signatures or lineage-associated polymorphisms.

From this base of six SNPs a group of eleven isolates also have non-synonymous SNPs in *rpoB* and a weakly annotated transcriptional regulator (*Gene_2267*). The product of *rpoB* is essential for transcription, its core function means it can act as a global regulator of gene function influencing all types of cell functions (Davati et al., 2023). The genome architecture and location of *Gene_2267* positions this gene upstream of a gene annotated as S-adenosylmethionine:tRNA ribosyltransferase-isomerase, commonly referred to as QueA. In *Salmonella*, enzymes in the queuosine biosynthetic pathway are critical for maintaining optimal growth and stress adaptation, which are essential for survival in diverse environments (Adeleye & Yadavalli, 2024). The changes observed among the 20EPA011NSA present an interesting snapshot of a *Salmonella* responding to a challenging environment.

3.5.2.4.3 20EPA012NSA – *S*. Infantis ST32

Two 20EPA012NSA isolates have a SNP in a gene named *bigA*. Not much is known about this gene, however based on homology the product of *bigA* likely functions as an autotransporter which could influence adhesion and biofilm formation during host-pathogen interactions (Curiao et al., 2016; Czibener et al., 2016).

3.5.2.4.4 22EPA051NSA – *S*. Enteritidis ST11

Two 22EPA051NSA isolates have a SNP in *dnaJ*. DnaJ is known to form chaperone machinery with DnaJ and GrpE with roles in many cellular processes, such as DNA replication, cell division, protein transport, RNA synthesis and autoregulation of the heat shock response. In mice the DnaK/DnaJ chaperone machinery has been shown to be essential for invasion of epithelial cells and survival within macrophages suggesting modification to DnaJ could be beneficial to survival in a human host (Takaya et al., 2004).

3.5.2.4.5 Pitfalls to the SNP analysis

One improvement to the study design would have been if I could plate stool on the day of collection from the diagnostic laboratory. This was not possible due to the potential of culturing a hazard group 3 *Salmonella*. Therefore, the stool specimen had to be stored at -80°C until it was referred to and sequenced by UKHSA, resulting in the identification of a serovar. This meant all samples were frozen before the culturing process. The time in storage varied dramatically as this project started before I had clearance to collect samples from the diagnostic laboratory. While I was receiving appropriate training, I was able to utilise *Salmonella* positive samples which had been collected by a colleague as negative controls for another project. The *S. Java* which I have seen the most genetic diversity at the SNP level had been in storage for over two years and this may have an influence on the results in this chapter. There is little direct evidence suggesting that *Salmonella* stored under appropriate conditions at -80°C accumulate significant SNPs over time. However, storing *Salmonella* in stool at an unknown stage in their growth cycle has the potential to cause an issue.

3.5.2.5 Genome Structure

The genome structure of all genomes screened within a stool sample was uniform and identified as GS1.0, aligning with GS1.0 being the most commonly observed genome structure across the *Salmonella* genus (Page et al., 2020). Structural deviations from GS1.0 are predominantly observed in *S.* Typhi and have been associated with its persistence within the human host (Page et al., 2020). A recent analysis of *S.* Agona isolates from UK infections identified GS1.0 as the most prevalent genome structure

(Waters et al., 2024). This study included sequencing isolates from both acute and persistent infections within individual patients, linking deviations from the GS1.0 structure to early convalescent carriage stages (Waters et al., 2024). The observation of GS1.0 in the isolates in this study supports that the *Salmonella* observed are generalists on temporary transit through the human gastrointestinal tract.

3.5.2.6 *S. Java* UK population

An intriguing observation is the disappearance of certain sequence types (STs) from circulation in the UK following the COVID-19 pandemic (Figure 3.10). Notably, ST42 appears to have been lost, along with ST135, ST423, and ST733. MLST and cgMLST analyses suggest that ST135, ST423, and ST733 likely represent expansions originating from ST42. Information on these STs is limited, although ST42 and ST423 have been mentioned in publications linking them to China and aquatic animals (Peng et al., 2024; Toboldt et al., 2013). While some STs have disappeared, others have remained stable or experienced temporary reductions in prevalence before showing signs of reestablishment. This time period of global population isolation could offer a chance to explore sequence types that are endemic to the UK versus those that are imported.

3.6 Conclusions

This study employed hybrid genome sequencing to investigate the genomic diversity of *Salmonella* isolates from a single patient's stool sample, uncovering insights relevant to pathogen characterisation and epidemiological analyses. Despite the challenges and variability associated with hybrid assemblies, the complementary strengths of long- and short-read sequencing were crucial for resolving structural genomic elements. The prevalence of a consistent GS1.0 among isolates is one of the main conclusions, indicating that these *Salmonella* were generalists temporarily circulating in the gastrointestinal system. Genetic diversity was observed in some isolates, with SNPs occurring in adaptive mechanisms, such as anaerobic survival and stress response, underscoring the potential for selective pressures to drive diversification. Significantly, there was little variance in AMR profiles, with the exception of one instance in which AMR genes were connected to mobile genetic components. This

emphasises the necessity of multi-isolate investigations in order to precisely capture genomic complexity and resistance mechanisms.

The study reinforces the need for improvements in sequencing technologies with approaches that balance resolution and cost. While single-colony studies are common practice they risk omitting significant genetic diversity. Emerging methods like sweep sequencing and metagenomics offer alternatives but come with their own limitations. Sequencing multiple isolates per sample provides the most reliable insights but remains resource intensive and expensive. This work emphasises the importance of tailoring genomic analyses to specific research and clinical contexts to enhance pathogen surveillance and public health responses. From an academic perspective, I believe a large-scale study replicating this work has the potential to identify genes under selective pressure during transit from farm to fork. In essence, this could serve as an assay to highlight potential targets for therapeutic intervention

4 Impact of Preservation Conditions on the Recovery of Metagenome Derived *Campylobacter* genomes from Stool Samples

4.1 Introduction

Campylobacter species are among the most common bacterial causes of GI illness worldwide(Kaakoush et al., 2015). They are notoriously fastidious organisms, often difficult to culture, and usually make up only a small proportion of the microbiome in diarrhoeal stool samples (B. Djeghout et al., 2024). In the United Kingdom, *Campylobacter jejuni* infections are the most commonly reported bacterial zoonosis, exceeding all other foodborne bacterial pathogens in incidence (Chlebicz & Slizewska, 2018; Kaakoush et al., 2015; Man, 2011). Traditional diagnostic methods for campylobacteriosis rely on culture, which can take 1–4 days and may miss viable but non-culturable cells (Khattak et al., 2022). Researchers are turning more and more to molecular techniques such as PCR and shotgun metagenomic sequencing, which offer quicker results and can pick up even tiny traces that might be missed by traditional methods. Analysing stool samples through direct sequencing is a powerful tool that can unlock valuable clinical insights. This approach lets us piece together complete genomes, helping identify specific bacterial strains, understand potential AMR resistance, and track different variants through genetic typing (Auguet et al., 2021; De, 2019; B. Djeghout et al., 2024; Peterson et al., 2022). This capability is especially promising as it bypasses the limitations of culture-based methods and allows for the reliable characterisation of a pathogen as fastidious as *Campylobacter*, which may otherwise go undetected in routine diagnostics (Mu et al., 2021; Peterson et al., 2022).

The rise of high-throughput sequencing has made shotgun metagenomic analysis a powerful tool in clinical microbiology (Harder et al., 2021). By sequencing all DNA in a stool sample, metagenomics can simultaneously detect multiple pathogens and characterise their genomes without the need for prior culture. Metagenomics is proving to be a valuable tool as we move away from traditional culture-based methods toward faster molecular testing for *Campylobacter*. While these newer sequencing techniques can give us a much fuller picture of the genomic landscape, their success

really comes down to having high-quality, representative DNA samples to work with (Fitzgerald et al., 2016). Reliable molecular analyses hinge on maintaining stable DNA levels through proper storage and workflow conditions, thereby avoiding post-sampling biases (Harder et al., 2021; O'Sullivan et al., 2018). In other words, we need to make sure our storage and shipping methods preserve the microbial community and pathogen DNA from when we first collected the sample. It's crucial we make the best effort to keep the sample as authentic as possible.

Stool samples present special challenges as a diagnostic specimen. Faeces is a heterogeneous matrix containing a complex community of gut microbes, shed host cells, digestive enzymes, and PCR inhibitors (Natarajan et al., 2021; Pereira-Marques et al., 2019). Even under ideal conditions, pathogenic *Campylobacter* DNA may represent only a tiny fraction of the total DNA pool (Dicksved et al., 2014; B. Djeghout et al., 2024). Furthermore, human DNA shed from intestinal cells can significantly contribute to background noise. When there is more host DNA in a sample, it becomes much harder to spot pathogen DNA using metagenomic sequencing (Pereira-Marques et al., 2019). Enzymes in stool (e.g. nucleases) and chemical factors can rapidly degrade nucleic acids if not properly inactivated. The combined consequence of these effects is that improper storage can diminish the production and fragment length of bacterial DNA, distorting the apparent community composition (Cardona et al., 2012; Granja-Salcedo et al., 2017; Panek et al., 2018). For example, analyses have shown that faecal microbiome profiles can change markedly after even 1–3 days at room temperature (Choo et al., 2015). Poor storage conditions and storage duration can make it harder to recovery *Campylobacter* by culture, and by extension would compromise DNA-based detection (Khattak et al., 2022).

DNA degradation, microbial composition shifts, and host DNA contamination can all impair both detection and typing of *Campylobacter*. Sequencing-based approaches require sufficient fragment lengths to cover marker genes for typing. While short amplicons (e.g. 100–300 bp) may remain detectable after days of storage (Harder et al., 2021), recovering full *Campylobacter* genomes or complete multilocus sequence types demands high-quality DNA. If storage causes substantial loss of pathogen DNA or increases human DNA, the accuracy of species detection, strain-typing, and AMR gene identification will suffer.

In clinical and epidemiological practice, several methods are used to preserve stool specimens. The simplest approach is raw freezing of stool aliquots at –80 °C (or colder) immediately after collection. This “flash-freezing” is widely regarded as the gold standard for nucleic acid preservation (Choo et al., 2015; Mehra & Kumar, 2024). Rapid freezing arrests enzymatic activity and microbial growth, helping maintain the sample’s original composition. However, freezing raw stool is believed to have practical drawbacks, it requires an uninterrupted refrigeration and can damage bacterial cell membranes when ice forms, potentially shearing DNA (Chen et al., 2022; Harder et al., 2021). An alternative is to use cryoprotective additives. Glycerol is commonly added to isolates and stool storage media to protect bacterial cells during freezing (Guerin-Danan, 1999; Li et al., 2023; Nursofiah et al., 2021). Early work showed that *C. jejuni* survival in stored samples was enhanced by glycerol-containing media (Gorman & Adley, 2004; Wasfy et al., 1995). Chemical stabilisers offer another strategy. Reagents such as RNAlater®, OMNIgene·GUT®, or DNA/RNA Shield (Zymo Research) are designed to inactivate microbes and preserve nucleic acids at ambient temperature. These are convenient for situations without immediate freezing capability. Little is known about the ability to store *Campylobacter* in stool using these chemical stabilisers. Notably, Zymo DNA/RNA Shield has been shown to preserve nucleic acids effectively in stool: one study reported significantly higher recovery of SARS-CoV-2 RNA from stool with DNA/RNA Shield than with Phosphate Buffered Saline (PBS) or alternative buffers (Natarajan et al., 2021). On the other hand, the immediate cell lysis caused by such reagents means that human DNA is released and captured along with microbial DNA, which can complicate metagenomic analyses by increasing host contamination (Bloomfield et al., 2023; T. Charalampous et al., 2019).

In this study I compare three representative conditions for preserving stool samples containing *Campylobacter* obtained from a clinical diagnostic laboratory: (1) raw stool aliquots frozen at -80°C (no additive); (2) stool diluted in glycerol-containing Brucella broth and frozen at -80°C; and (3) stool mixed with DNA/RNA Shield (a proprietary nucleic acid stabiliser) and stored frozen at -80°C. These methods were chosen because they represent common practices and commercially available options in clinical and research laboratories.

4.2 Aims and objectives

The work outlined in this chapter aimed to assess a set of preservation methods that optimally preserve *Campylobacter* DNA for metagenomic sequencing, thereby ensuring accurate pathogen detection and typing in diagnostic and surveillance applications as outlined in the aims below:

- Establish a benchmark collection of *Campylobacter* genomes for comparison with metagenomic results by isolating *Campylobacter* from each stool sample prior to storage.
- Extract DNA from stool samples prior to storage as a baseline sample comparison.
- Store stool in three sample preservation conditions: raw stool (no preservative), in broth with glycerol, and in Zymo DNA/RNA Shield.
- Extract DNA from stool stored in three preservation conditions at -80°C after 1, 3, and 9 months
- Sequence metagenomes of stool samples and compare metrics including classification, sequence type, AMR profile, and genome coverage.
- Perform quantitative polymerase chain reaction (qPCR) to quantify both *Campylobacter* and human DNA loads.

4.3 Methods

4.3.1 Experimental design overview

Stool samples were stored under three conditions: raw (no preservative), in broth with glycerol, and in Zymo DNA/RNA Shield. DNA was first extracted from the stool on the day of collection, and subsequently from each preservation condition stored at -80°C after 1, 3, and 9 months. At the time of collection, *Campylobacter* was cultured and sequenced from each sample (Figure 4.1). Isolate genome data were used to generate a *Campylobacter* genotype profile for each stool sample, including species identification, sequence type, and antimicrobial resistance determinants. This profile was then screened for in each metagenome derived (MD)-*Campylobacter* genome from the respective stool sample. A single isolate from each stool sample was selected as a reference, and metagenomic reads from each condition and time point were mapped to this reference to generate coverage scores. Quast (Galaxy v 5.0.2) was employed to assess genome completeness metrics of isolate assemblies, selecting a reference for each stool based on criteria including the highest N50 value, lowest number of contigs, and largest contig size. In addition, qPCR was performed on all DNA preparations to detect and quantify the presence of *Campylobacter* and human DNA.

4.3.2 Sample collection

Surplus diarrhoeal stool specimens were collected from the National Health Services Eastern Pathology Alliance (EPA) network diagnostic laboratory, Norwich, Norfolk, UK. Stool specimens represented four separate anonymised patients with gastroenteritis symptoms who submitted specimens to the laboratory between June 2023 and July 2024. *Campylobacter* spp. were initially identified in the stool specimens by the diagnostic laboratory using a rapid automated PCR-based culture-independent testing panel (Gastro Panel 2, EntericBio, Serosep United Kingdom). Once PCR results were confirmed, a 15-20 mL aliquot of stool was placed into a sterile specimen container and transported to Quadram Institute Bioscience in a triple-contained container.

In total, twelve stool specimens were processed in this study. The number of samples reflected a balance between the financial and logistical constraints of performing both

isolate-level and metagenomic sequencing at multiple storage time points. Additionally, the time required to obtain stool specimens of sufficient volume and that yielded *Campylobacter* colonies further limited the total sample size, while still ensuring meaningful comparative analyses across treatments within the project time window.

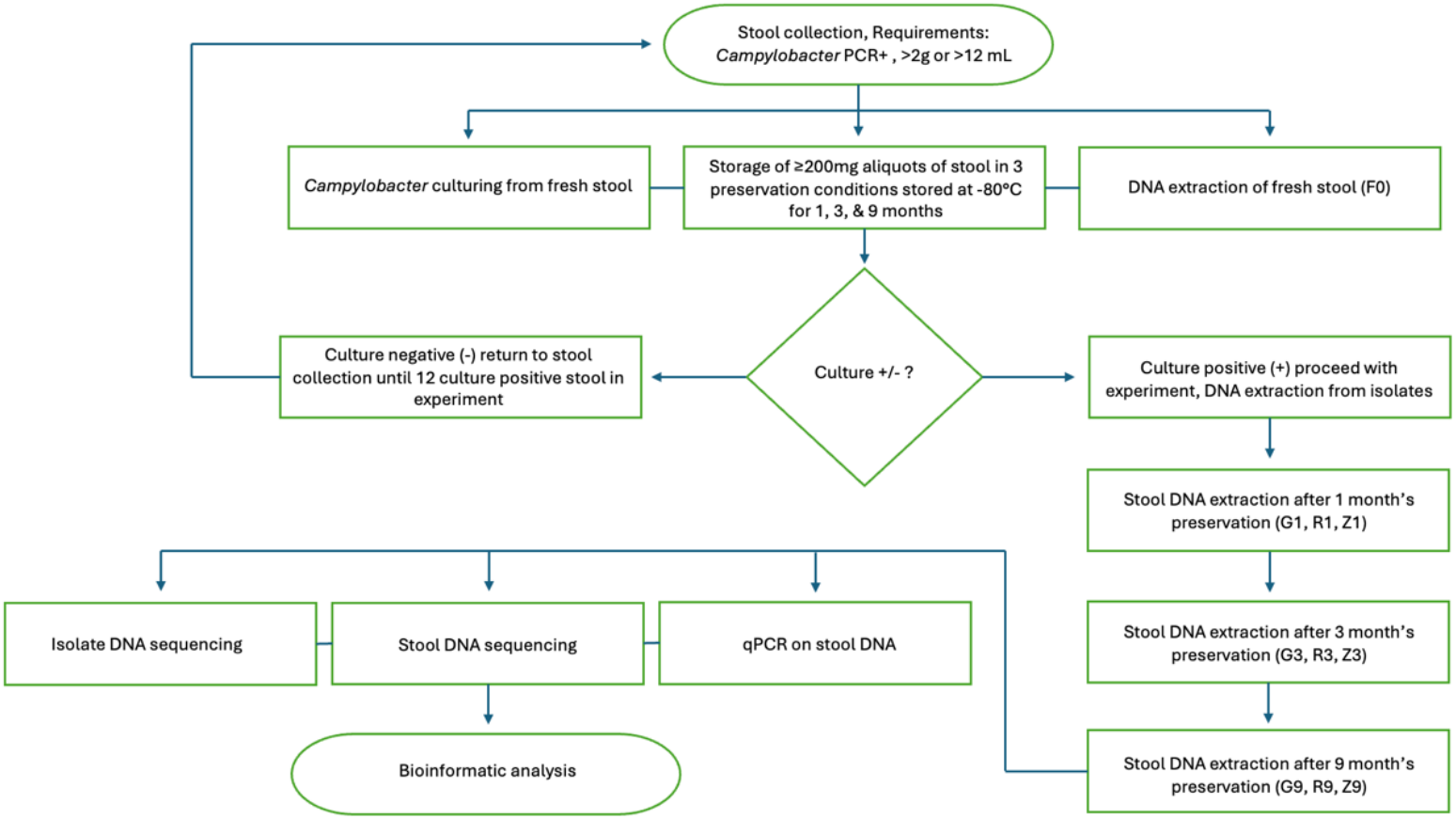

Figure 4.1: Flowchart of experimental design

4.3.3 Bacterial isolation

Stool was plated on *Campylobacter* Blood-Free Selective Medium (modified Charcoal Cefoperazone Deoxycholate Agar (mCCDA)). The media was prepared by QIB core laboratory support technical staff. A 10 µL aliquot of stool was streaked onto mCCDA plates and incubated in a microaerophilic atmosphere using anaerobic jars with a CampyGen 2.5 L sachet (Oxoid, Hampshire, UK) at 37 °C for 48 h. Putative *Campylobacter* colonies were streaked onto Columbia Blood Agar (CBA) and incubated in a microaerophilic atmosphere using anaerobic jars with a CampyGen 2.5 L sachet at 37 °C for 48 hours. An oxidase test was used to screen colonies cultured based on visual colony morphology identification. In brief, a vial of Remel BactiDrop (Remel Inc. (Thermo Fisher Scientific), USA) was poured onto sterile Whatman filter paper (Whatman International Ltd., UK) within a petri dish. Using a plastic loop, a small amount of bacterial material was collected and spread onto the BactiDrop soaked paper. A change in colour to purple was interpreted as an oxidase-positive result, and those colonies were prepared for sequencing using the Maxwell RSC Fecal Microbiome DNA Kit (Promega, USA). The remainder of the isolate was preserved in broth + 15% glycerol and stored at -80 °C. A minimum of 6 isolates was collected per stool samples with the maximum being 12.

4.3.4 Stool sample preservation conditions and storage

Upon collection, 200 mg aliquots of stool were stored in 2 mL Cryo vials (ref 202035-1, Altemis Lab, UK) as raw, in a 200 µL Brucella broth with 17.5 % glycerol, and in 500 µL 2x Zymo DNA/RNA shield. A minimum of three aliquot per condition were stored at -80 °C. For samples that were Bristol scale 6 or 7 (watery diarrhoea) a 200 mg pellet was collected by centrifuging the stool sample. The supernatant was removed and excluded from the 200 mg stool sample.

4.3.5 DNA extraction from stool

The stool samples were prepared by placing approximately 200 mg or 200 µL of raw stool into labelled 2 mL round bottom sterile Eppendorf tubes, using either a sterile 10

µL loop or a 1000 µL pipette. For diarrheal samples (Bristol scale 7), 1-2 mL replicates were centrifuged to obtain a solid pellet suitable for Host Depletion (HD) treatment, ensuring each pellet weighed 200 mg. Then, 200 µL of HD-buffer was added to each tube containing the stool sample, taking care not to exceed 300 mg of stool during this step. Next, 10 µL of HL-SAN enzyme was added to each tube containing the HD-buffered stool, and each sample was gently vortexed for 30 seconds to ensure thorough mixing. The samples were incubated at 37 °C for 20 minutes using an Eppendorf shaking heat block or a HulaMixer in a 37 °C incubator. Once the incubation was complete, each tube was vortexed again to ensure thorough mixing of the HD-treated samples. The tubes were then centrifuged at 10,000 rcf for 5 minutes to pellet the HD-treated samples, and the supernatant was carefully removed. The resulting HD pellet represented the lysed stool sample for further analysis.

1 mL of Lysis Buffer and 40 µL of Proteinase K were added to the microcentrifuge tube containing the HD pellet, which was then vortexed for 30 seconds. The tube was placed into a heat block at 95°C for 5 minutes, after which the samples were removed and allowed to cool for 2 minutes on the benchtop. The samples were vortexed thoroughly for 1 minute, followed by an incubation at 56°C for 5 minutes. During this incubation, cartridges were prepared as outlined in the 'Preparing the Cartridge' section on the next page. The lysate tubes were centrifuged in a microcentrifuge at room temperature for 5 minutes at maximum speed (10,000g rcf) to pellet solids. Finally, 300 µL of supernatant was transferred into well #1 of the reagent cartridges, ensuring to avoid pipetting any solid material from the tube bottom or oil from the liquid surface. If necessary, the supernatant was transferred to a new tube and centrifuged again to remove any remaining solids.

The cartridges were placed in the deck tray(s) with well #1 (the largest well) facing away from the elution tube. Each cartridge was snapped into position, and the seal from the top was carefully removed to ensure all sealing tape and residual adhesive were cleared before placement in the instrument. Cartridges were handled with care, noting sharp seal edges. A plunger was inserted into well #8 of each cartridge (closest to the Elution Tube). An empty Elution Tube was placed into the Elution Tube position for each cartridge in the deck tray(s), ensuring that caps were open and facing away from the cartridge positions. 100 µL of Elution Buffer was added to the bottom of each Elution Tube. It was noted that optimal elution may be compromised if Elution Buffer is

on the side of the tube, so only the provided 0.5 mL Elution Tubes were used as other tubes may not be compatible with the Maxwell RSC Instrument. Additionally, 300 µL of Binding Buffer was added to well #1 of each cartridge, and 20 µL of RNase A was added to well #3 of each cartridge. The setup and run instructions detailed in the Maxwell RSC Fecal Microbiome DNA Kit Technical Manual #TM640 were followed.

4.3.6 DNA extraction from isolates

Campylobacter isolates were recovered from stool samples by culture, as described in section 4.3.4. A 10 µl aliquot of stool was streaked onto mCCDA and plates and incubated in a microaerophilic atmosphere using anaerobic jars with a CampyGen 2.5 L sachet (Oxoid, Hampshire, UK) at 37 °C for 48 h. *C. jejuni* strain 81116 was used as a positive control for growth. Cells were collected from the plate with a plastic loop and resuspended in PBS. The cells were collected as a pellet by centrifuging and used as input into the *Salmonella* Fire Monkey HMW DNA extraction protocol described in 2.4.2.2. DNA was stored at -20 °C in two aliquots to avoid freeze-thaw cycles, one for sequencing and one for qPCR.

4.3.7 DNA quantification

The Qubit dsDNA BR Assay Kit (Q32853, Thermo Fisher, UK) was used to quantify DNA prior to DNA sequencing (Ref. Q32853, Thermo Fisher Scientific, USA). The working solution was prepared by adding 199 µL of Qubit dsDNA BR buffer and 1 µL of Qubit dsDNA BR dye, this was made as a master mix for the desired number of samples. To prepare standards, 190 µL of working solution and 10 µL standard were mixed in a Qubit assay tube (Ref. Q32856, Thermo Fisher Scientific, USA). For samples, 198 µL of working solution was mixed with 2 µL DNA. All tubes were incubated at room temperature for 2 minutes to ensure proper binding of the dye to DNA molecules. The Qubit 3.0 fluorometer was set up, calibrated using the standards, and the fluorescence of each standard and sample was measured using the Qubit dsDNA BR assay program.

4.3.8 Quantitative qPCR

All qPCR assays were performed in triplicate on a Roche LightCycler 480 II using LightCycler 480 SW 1.5 for analysis. A 4-colour hydrolysis probe design was utilised with Abs Quant / 2nd derivative max analysis using a comb filter set at 498-580 nm. The cycles were as follows: 1 cycle pre-amplification at 95°C for 10 minutes with 4.4 °C/s ramp rate, 45 cycles amplification at 95°C for 0:15 seconds with 4.4 °C/s ramp rate, followed by 55°C for 1:00 minute with 2.2 °C/s ramp rate. The protocol was finished with 1 cooling cycle at 40°C for 0:30 seconds with 2.2 °C/s ramp rate.

The *cadF* gene assay reactions used to identify *Campylobacter* were set up as follows: 10 µL LightCycler 480 Probe Mix (Cat. Number 04707494001, Roche, Switzerland), 0.4 µL *cadF* forward (10 µM), 0.4 µL *cadF* reverse (10 µM), 0.2 µL *cadF* probe (10 µM), 7 µL H_20 (supplied with the LightCycler 480 Probe Mix), and 2 µL of sample DNA at 10 ng/µL. Genome DNA of *C. jejuni* strain 13361 was used as a positive control.

The human assay reactions used to identify the presences of human DNA were set up as follows: 10 µL LightCycler 480 Probe Mix (Cat. Number 04707494001, Roche, Switzerland), 0.5 µL Human forward (10µM), 0.5 µL Human reverse (10µM), 0.4 µL Human probe (10µM), 6.6 µL H_20 (supplied with the LightCycler 480 Probe Mix), and 2 µL of sample DNA at 10 ng/µL. TaqMan Control Human Genomic DNA (Cat. 4312660, Thermo Fisher, USA) was used as a positive control.

4.3.9 DNA sequencing library preparation

Genomic DNA was normalised to 5 ng/µL with EB (10 mM Tris-HCl), this process was the same for bacterial isolates and metagenome libraries. A master mix was prepared by combining 0.5 µL of Tagmentation Buffer with 0.5 µL of Bead Linked Transposomes (Illumina Catalogue No. 20018704) and 4 µL of PCR grade water. 5 µL of this tagmentation mix was added to each well of a 96-well plate. Next, 2 µL of normalised DNA (10 ng total) was mixed with the tagmentation mix in each well and heated to 55°C for 15 minutes in a PCR machine. For PCR amplification, a master mix was prepared using 10 µL of KAPA 2G Fast Hot Start Ready Mix (Merck Catalogue No. KK5601) and 2 µL of PCR grade water per sample. 12 µL of this master mix was added to each well of

the 96-well plate. Additionally, 1 µL of a 10 µM primer mix containing both P7 and P5 Illumina 9 bp barcodes was added to each well. The final step involved adding 7 µL of the tagmentation mix to each well and thoroughly mixing. PCR cycling conditions were set as follows: initial denaturation at 72°C for 3 minutes, followed by 14 cycles of 95°C for 10 seconds, 55°C for 20 seconds, and 72°C for 3 minutes. After PCR, libraries were quantified using the Promega QuantiFluor® dsDNA System (Catalogue No. E2670) and measured on a GloMax® Discover Microplate Reader. Libraries were pooled in equal quantities and subjected to double-SPRI size selection between 0.5 and 0.7X bead volumes using sample purification beads (Illumina® DNA Prep, (M) Tagmentation (96 Samples, IPB), 20060059). The final library pool was quantified using a Qubit 3.0 instrument and analysed on an Agilent Tapestation 4200 using a D5000 ScreenTape (Agilent Catalogue No. 5067-5579) to determine the final library pool molarity.

4.3.10 DNA sequencing

Bacterial isolate sequencing using Illumina paired-end 150bp was carried out by QIB sequencing on a NextSeq 500. The pool, adjusted to a final concentration of 1.5 pM, was sequenced on an Illumina Nextseq500 instrument using a Mid Output Flowcell (NSQ® 500 Mid Output KT v2 (300 cycle), Illumina Catalogue FC-404-2003), following Illumina's recommended denaturation and loading protocols, which included a 1% PhiX spike in (PhiX Control v3, Illumina Catalogue FC-110-3001). Metagenome sequencing was Illumina paired-end 150bp, carried out externally by Novogene and Azenta on an Illumina Novaseq X.

All metagenomic sequencing was outsourced to external service providers, library prep was completed by QIB sequencing as described in 4.2.9. Initially, ten samples were sequenced alongside another project by Novogene. This was followed by a batch of 96 samples and one blank sent to Novogene, and a final batch of 52 samples plus a blank sent to Azenta. The final sequencing run included some repeat samples that had failed to reach the target yield of 8 Gb in an earlier run. Each blank consisted of 200 µL of PCR-grade water processed using the protocol described in Section 4.2.4. After completion of the three sequencing runs a single sample failed to reach the 8Gb target (141_G_9M, 4.97Gb). Sample 132_F0 was included on all three sequencing runs. Full lists of sequencing statistics can be found in Appendix 2.

4.3.11 *in-silico* human read removal

Human host DNA depletion was conducted at the sample preparation stage of stool metagenomes; however, additionally, it is standard practice at Quadram to run *in-silico* human read removal before uploading data to the QIB instance of IRIDA, a genome storage platform. This process was carried out by the QIB core bioinformatics team using Centrifuge (Galaxy v1.0.3) with the database human-t2t-hla.argos-bacteria-985_rs-viral-202401_ml-phag (Kim et al., 2016).

4.3.12 Bacterial isolate assembly

Paired-end read files were first processed with fastp (Galaxy v0.23.2) using default settings. The processed reads were fed into Shovill (Galaxy v1.0.4) for assembly with SPAdes.

4.3.13 *Campylobacter* read recovery from metagenome sequencing

Metagenomic sequencing files were first processed with fastp (Galaxy v0.23.2) using default settings. Processed sequencing files were assigned taxonomic labels using Kraken2 (Galaxy v2.1.3) with confidence set at 0.2. The selected database was k2_nt_20230502. *Campylobacter* reads were extracted using Krakentools (Galaxy version 1.2) using the taxonomic ID 194.

4.3.14 Recovered read assembly

The read recovery process led to irregular pair-end sets so forward and reverse read files were collapsed into single read files within Galaxy (v4.2). The reads were assembled using Megahit (Galaxy v1.2.9). Metagenome-derived *Campylobacter* genomes are identified as MD-*Campylobacter* genomes.

4.3.15 Classification

Contigs from bacterial isolate sequencing and MD-*Campylobacter* genomes were classified using GTDB-Tk (Galaxy v2.2.2) with database gtdb-20190917.

4.3.16 Multi-Locus Sequencing Typing

Multi-Locus Sequence Typing (MLST) was performed on contigs from *Campylobacter* isolate sequencing and the MD-*Campylobacter* genomes scheme using MLST (Galaxy v2.16.1) using parameters set at 95% for minimum DNA % identity and 10% for minimum DNA % coverage. For samples where an allele profile was obtained with MLST but not an overall ST number the website https://pubmlst.org/ was used to navigating to *Campylobacter jejuni*/*coli* typing. The allele numbers from MLST were manually entered into the website form.

4.3.17 Antimicrobial resistance genotyping

Antimicrobial resistance determinants of isolate-derived and MD-*Campylobacter* genomes were identified using abriTAMR (Galaxy v1.0.14) set to detect *Campylobacter*-specific point mutation acquired resistance and resistance gene presence. The program was run in default settings which sets minimum identity of matches with armfinder to 0.9.

4.3.18 Read mapping

Read mapping was carried out on a MacBook Pro (Apple M1) running macOS Sonoma. To calculate coverage, score for breadth and depth a reference bacterial isolate for each stool sample was indexed using bwa (v0.7.18) using the code “bwa index isolate.fasta”. Next the paired-end reads were aligned to the consensus genome using the code “bwa mem isolate.fasta forward.fastq.gz reverse.fastq.gz > paired_reads_vs_consensus.sam”. Next the SAM file was converted to a BAM file using the code “samtools view -b paired_reads_vs_consensus.sam | samtools sort -o paired_reads_vs_consensus.sorted.bam”. The BAM file was indexed using the code “samtools index paired_reads_vs_consensus.sorted.bam”. A general coverage score was calculated with the code “samtools depth paired_reads_vs_consensus.sorted.bam > coverage.txt”. This was summarised across the genome using the code “awk '{sum+=$3; count++} END {print "Average Coverage: ", sum/count}' coverage.txt”. Breadth of coverage was calculated using samtools (1.16.1) using the code “samtools depth -a align.sorted.bam | awk '{if($3>0) count++} END {print count/GenomeLength*100}'”. GenomeLength for the reference genomes were obtained from Quast (Galaxy v5.0.2) and was calculated from the contig output

from Shovil after assembly (Section 3.2.9). Depth of coverage was calculated with the code “samtools depth align.sorted.bam | awk '{sum+=$3} END {print sum/GenomeLength}'”.

4.3.19 Genome assembly quality assessment by QUAST

All assembled isolate genomes were assessed using QUAST (Galaxy v5.0.2). QUAST statistics including Total Length, N50, GC (%), and # contigs were used to select a reference for each stool sample. MD-*Campylobacter* genomes were run through QUAST using the selected references as a reference genome to obtain a N50 score and a genome fraction score for further statistical analysis (Gurevich et al., 2013).

4.3.20 CheckM analysis of metagenome derived genome (MDG) completeness

MD-*Campylobacter* genomes were assessed with CheckM (Galaxy v1.2.0) with taxonomic rank set to genus, and taxon of interest set to *Campylobacter*.

4.3.21 Statistics

Statistical tests were carried out using python in JupyterLab v4.2.1 launched through Anaconda Navigator v2.6.0. The Shapiro-Wilk test was imported from scipy.stats. The Shapiro–Wilk test was used to assess the normality of data distributions. This test evaluates whether a dataset follows a normal (Gaussian) distribution, where $p > 0.05$ indicates no significant deviation from normality. As most datasets violated the assumption of normality ($p < 0.05$), non-parametric tests were applied.

For the Wilcoxon test, the basic test was imported from scipy.stats. The Benjamini-Hochberg false discovery rate correction applied to the Wilcoxon test was imported form statsmodels.stats.multitest. This test evaluates whether the median difference between paired observations differs significantly from zero, providing a robust alternative to the paired t-test when data are not normally distributed. To account for multiple comparisons, p-values obtained from the Wilcoxon tests were adjusted using the Benjamini–Hochberg (BH) false discovery rate (FDR) correction.

The logistic regression model was imported from statsmodels.Logit. A multivariable logistic regression model was fitted to evaluate which factors were associated with successful sequence type (ST) assignment, defined as an ST Score of 7 (ST7 = 1). The outcome variable was binary: samples that achieved an ST Score of 7 were coded as 1, and all others (ST Score < 7) were coded as 0. The logistic regression model included several predictor variables to assess their influence on the likelihood of achieving ST Score = 7. These predictors were: *cadF* mean Cp (continuous), representing the qPCR crossing point (Cp) value for the Campylobacter *cadF* gene as a proxy for bacterial load; Human mean Cp (continuous), indicating the Cp value for human DNA and serving as a proxy for host DNA abundance; Condition (categorical), representing the storage condition of each sample (F, G, R, or Z), with Condition F used as the reference group; and Timepoint (continuous), representing the storage duration in months (0, 1, 3, or 9). The model estimated the effect of each predictor on the log-odds of achieving ST Score = 7, adjusting for the influence of all other variables in the model.

Packages matplotlib.pyplot as plt and seaborn as sns were used to create plots. The package Pandas was used throughout to enable processing of excel and csv files. The NumPy package was loaded as standard practise to support arrays and numerical function. Two approaches were used to normalise the metagenomic coverage metrics (Breadth, Depth, and Genome fraction). First, values were normalised by sequencing depth by dividing each metric by the total number of reads (reads_in) and scaling to reflect coverage per 10 million reads. This allowed for direct comparison across samples with differing sequencing depths. Second, a $\log_{10}$ transformation was applied. Metrics were first normalised by reads_in, and then transformed using the base-10 logarithm. To accommodate zero values in the dataset and avoid undefined log operations, a small pseudo count ($1e^{-6}$) was added prior to transformation.

4.4 Results

4.4.1 Overview of stool samples

In total, 12 rapid PCR *Campylobacter*-positive stool specimens were collected between July 2023 and March 2024. Despite *Campylobacter* being consistently present at the EPA laboratory, it required repeated weekly visits to collect 12 samples

that met the study inclusion requirements. A collection window was established, and stool specimens were collected only within three days of their receipt at the PA laboratory (Table 4.1). This window was selected as predicted to be the shortest time frame in which sufficient samples could be collected during this project. All stool samples spent 1-3 days at the EPA laboratory undergoing testing and an unknown time period making their way to the EPA laboratory from the hospital or community setting (e.g. general practices). The volume collected for 23EPA130C and 23EPA135C was sufficient to allow for two aliquots of stool to be stored in each condition for DNA extraction at each timepoint.

Table 4.1: Summary of Stool Samples in the Storage Conditions Experiment

Full Sample ID	Short ID	Date QIB collection	Date of submission to EPA	Days in fridge at EPA
23EPA124C	124	06/07/2023	04/07/2023	2
23EPA128C	128	14/07/2023	11/07/2023	3
23EPA130C	130	20/07/2023	18/07/2023	1
23EPA132C	132	27/07/2023	24/07/2023	3
23EPA135C	135	24/08/2023	22/08/2023	2
23EPA136C	136	24/08/2023	22/08/2023	2
24EPA141C	141	11/01/2024	09/01/2024	2
24EPA143C	143	18/01/2024	16/01/2024	2
24EPA144C	144	08/02/2024	07/02/2024	1
24EPA145C	145	08/02/2024	05/02/2024	3
24EPA146C	146	08/02/2024	05/02/2024	3
24EPA147C	147	07/03/2024	06/03/2024	1

Labelling conventions used throughout the results section are as follows: Stool samples are identified by a shortened version of their full name, for example 23EPA124C is referred to simply as 124. Timepoints 1, 3, and 9 correspond to months after the start of the experiment. F0 refers to sequencing performed on a stool sample prior to storage. Storage conditions are abbreviated as follows: R for raw stool (no preservation), G for stool stored in Brucella broth with 17.5% glycerol, and Z for stool stored in a 5:1 ratio of Zymo DNA/RNA Shield to stool. Occasionally, particularly in plots, samples are labelled with combinations such as R1, indicating raw-stored stool sampled at 1 month. Two stools were included as biological replicates, and these are identified with R1 and R2 after the stool_id (e.g. 135R1 and 135R2).

4.4.2 *Campylobacter* isolation

Each stool sample was processed to isolate *Campylobacter*, with up to 12 colonies selected for subsequent storage and sequencing. At least six *Campylobacter* isolates were obtained for all samples except for 23EPA128C. Despite repeated attempts, efforts to culture *Campylobacter* from this sample were unsuccessful, as the culture was consistently dominated by competing bacteria. In total, six isolates were cultured and sequenced from stool 23EPA128C. Five of these isolates were *Ochrobactrum anthropic* and one was unclassified by GTDB-Tk. Among the remaining samples, *C. jejuni* was exclusively identified in ten instances, while *C. coli* was the sole species detected in one stool sample. At the ST level, a single ST was observed from all isolates within each stool sample, with a different ST observed for each stool sample. At the clonal complex level, the most commonly observed complexes were CC-21 and CC-353 (Table 4.2).

Table 4.2: Classification, Sequence Type (ST), and AMR Determinants Identified in *Campylobacter* Isolates from Stool Samples

Stool ID	Classification	ST	CC	No. of isolates	AMR profile (number of isolates with AMR determinant)
124	*Campylobacter jejuni*	464	464	10	*bla*$_{OXA-193}$ (10), *tet(O)* (10), *gyrA*_T86I (10), *50S*_L22_A103V (10)
128	Undetermined	n/a	n/a	0	n/a
130	*Campylobacter jejuni*	791	Singleton	9	*bla*$_{OXA-184}$ (9), *tet(O)* (8)
132	*Campylobacter jejuni*	10846	353	12	*bla*$_{OXA-193}$ (12), *tet(O)* (12), *gyrA*_T86I (12), *50S*_L22_A103V (12)
135	*Campylobacter jejuni*	1707	607	12	*bla*$_{OXA-193}$ (12), *tet(O)* (12)
136	*Campylobacter jejuni*	4697	353	6	*50S*_L22_A103V (6)
141	*Campylobacter jejuni*	9897	Singleton	12	*bla*$_{OXA-193}$ (12), *tet(O)* (12), *gyrA_T86I* (12)
143	*Campylobacter jejuni*	21	21	12	*bla*$_{OXA-193}$ (12), *gyrA_T86I* (12)
144	*Campylobacter jejuni*	6175	21	12	*bla*$_{OXA-193}$ (12), *tet(O* (12), *L* (4), *M* (4)), *cepA* (1)
145	*Campylobacter coli*	829	828	11	*bla*$_{OXA-489}$ (11), *tet(O* (11), *L*(2), *Q* (4), *X1*(2), *X2*(2)), *gyrA_T86I* (11), *cfxA*(8), *lnu(C)* (2), *dfrF* (2), *bexA* (2), *aadS* (2)
146	*Campylobacter jejuni*	19	21	9	*bla*$_{OXA-193}$(9)
147	*Campylobacter jejuni*	400	353	12	*tet(O)* (12), *gyrA*_T86I (12), *50S*_L22_A103V (12)

n/a = no data due to stool sample being culture negative, ST = sequence type, CC = clonal complex

The genotypic AMR profiles observed in the *Campylobacter* isolates from the stool samples were diverse and include several key determinants (Table 4.2). Specifically, *blaOXA-193* was prominently present across multiple *C. jejuni* isolates from various stool samples (samples 124, 132, 135, 141, 143, 144, 146, 147), contributing primarily to resistance against beta-lactam antibiotics. The gene *tet*(O) was identified in samples 124, 130, 132, 135, 141, 144, and 145, which confers resistance to tetracyclines. The *gyrA*_T86I mutation, observed in samples 124, 132, 143, and 147, is associated with resistance to fluoroquinolones. The *50S*_L22_A103V mutation, found in samples 124, 132, 135, and 147, contributes to resistance against macrolide antibiotics. Other AMR determinants include *blaOXA-184* in sample 130, *tet(L)* and *tet(M)* in sample 144 (tetracycline resistance), and *blaOXA-489*, *tet(Q)*, *tet(X1)*, and *tet(X2)* in sample 145 (resistance to beta-lactams and tetracyclines). Additionally, resistance to other antibiotics such as cephalosporins (*cfxA*), aminoglycosides (*Inu*(*C*), *aadS*), and trimethoprim (*dfrF*) was observed in sample 145 (Table 4.2).

4.4.3 Sequence typing for metagenome derived genomes

MD-*Campylobacter* genomes from all stool samples, conditions and timepoints were screened to obtain sequence type information. This information was then compared to the sequence type information obtained from the isolates recovered from the same stool samples. For samples sequenced from fresh stool (F0), the F0 MD-*Campylobacter* genomes returned a complete ST for 6 out of 12, that is, all 7 alleles in the scheme were correct. These were for stool IDs 124, 132, 135, 141, 143, and 147. F0 MD-*Campylobacter* genomes for stool IDs 136 and 146 matched 6 alleles correctly, and from 3 stool IDs 0 alleles matched correctly, these were 130, 144, and 145. When 0 alleles were matched at F0 this remained the cases throughout the storage experiment. For stool IDs 132 and 135, the count remained at 7 throughout the experiment, indicating that full sequence type information was successfully recovered across all storage conditions and timepoints. For stool ID 143, no sequence type information was recovered from condition Z at any timepoint. In contrast, complete ST profiles were obtained for conditions G and R. Varying patterns of sequence type information were recovered from stool IDs 124, 136, 141, 146, 147. For these samples, preservation conditions R and G performed better than preservation condition Z (Table

4.3). Overall, MD-*Campylobacter* genomes in preservation condition Z showed the lowest DNA quality, whereas DNA from conditions R and G was of higher and comparable quality, with R slightly higher. For 128, no isolates were cultured from the stool, and no ST information was recovered by sequencing. No differences were observed with the biological replicates R1s and R2s.

Table 4.3: MLST Allele Score for Metagenome Derived Genomes Versus Isolate Reference Genome, 7 Alleles Represent a Complete MLST Profile Resulting in a Sequence Type

	MLST allele score at each preservation condition and time point									
MDG ID	**F0**	**G1**	**G3**	**G9**	**R1**	**R3**	**R9**	**Z1**	**Z3**	**Z9**
124	7	3	7	6	0	1	1	1	6	1
128	0	0	0	0	0	0	0	0	0	0
130R1	0	0	0	0	0	0	0	0	0	0
130R2	0	0	0	0	0	0	0	0	0	0
132	7	7	7	7	7	7	7	7	7	7
135R1	7	7	7	7	7	7	7	7	7	7
135R2	7	7	7	7	7	7	7	7	7	7
136	6	3	4	4	3	0	4	0	0	0
141	7	0	1	0	6	0	7	0	0	0
143	7	7	7	7	7	7	7	0	0	0
144	0	0	0	0	0	0	0	0	0	0
145	0	0	0	0	0	0	0	0	0	0
146	6	4	3	2	2	0	1	0	0	0
147	7	2	6	0	3	4	6	0	0	0

F0 = DNA extracted from fresh stool at Time 0 (at time of collection). Preservation conditions: G = Stool stored in broth with glycerol; R = Raw stool (no preservation); Z = Stool stored in Zymo DNA/RNA shield. Numbers indicate storage duration at -80°C: 1 = 1 month, 3 = 3 months, and 9 = 9 months. MDG = Metagenome-derived genome.

4.4.4 Classification

No classification was obtained for sample 128, suggesting that if *Campylobacter* cannot be cultured from a PCR-positive stool sample, MDG sequence-based identification is also likely to be unsuccessful. For F0 MDGs, 8 out of 12 returned the correct *Campylobacter* species. The results mirror the sequence typing pattern, showing minimal difference between conditions R and G, both of which produced more correct classifications than condition Z. Some incorrect classifications were identified in samples 124 R1, 124 R3, 124 Z1, and 136 Z3 (Table 4.4).

Table 4.4: GTDB-Tk Classification of Metagenome Derived Genomes Compared to Isolate References

	***Campylobacter* classification at each preservation condition and time point**									
Stool (Isolate Classification)	**F0**	**G1**	**G3**	**G9**	**R1**	**R3**	**R9**	**Z1**	**Z3**	**Z9**
124 (*C. jejuni*)	*C. jejuni*	*Campylobacter*	*C. jejuni*	*C. jejuni*	*C. coli*	*C. hepaticus*	*Campylobacter*	*C. coli*	*C. jejuni*	*C. jejuni*
128 (n/a)	Unclassified	Unclassified	Unclassified	Unclassified	Unclassified	Unclassified	Unclassified	Unclassified	Unclassified	Unclassified
130R1 (*C. jejuni*)	Unclassified	Unclassified	Unclassified	Unclassified	Unclassified	Unclassified	Unclassified	Unclassified	Unclassified	Unclassified
130R2 (*C. jejuni*)	Unclassified	Unclassified	Unclassified	Unclassified	Unclassified	Unclassified	Unclassified	Unclassified	Unclassified	Unclassified
132 (*C. jejuni*)	*C. jejuni*	*C. jejuni*	*C. jejuni*	*C. jejuni*	*C. jejuni*	*C. jejuni*	*C. jejuni*	*C. jejuni*	*C. jejuni*	*C. jejuni*
135R1 (*C. jejuni*)	*C. jejuni*	*C. jejuni*	*C. jejuni*	*C. jejuni*	*C. jejuni*	*C. jejuni*	*C. jejuni*	*C. jejuni*	*C. jejuni*	*C. jejuni*
135R2 (*C. jejuni*)	*C. jejuni*	*C. jejuni*	*C. jejuni*	*C. jejuni*	*C. jejuni*	*C. jejuni*	*C. jejuni*	*C. jejuni*	*C. jejuni*	*C. jejuni*
136 (*C. jejuni*)	*C. jejuni*	*C. jejuni*	*C. jejuni*	*C. jejuni*	*C. jejuni*	*C. jejuni*	*C. jejuni*	*C. jejuni*	*C. coli*	Unclassified
141 (*C. jejuni*)	*C. jejuni*	Unclassified	Unclassified	Unclassified	*C. jejuni*	*C. jejuni*	*C. jejuni*	Unclassified	Unclassified	*Campylobacter*
143 (*C. jejuni*)	*C. jejuni*	*C. jejuni*	*C. jejuni*	*C. jejuni*	*C. jejuni*	*C. jejuni*	*C. jejuni*	Unclassified	Unclassified	Unclassified
144 (*C. jejuni*)	Unclassified	Unclassified	Unclassified	Unclassified	Unclassified	Unclassified	Unclassified	Unclassified	Unclassified	Unclassified
145 (*C. coli*)	Unclassified	Unclassified	Unclassified	Unclassified	Unclassified	Unclassified	Unclassified	Unclassified	Unclassified	Unclassified
146 (*C. jejuni*)	*C. jejuni*	*C. jejuni*	*C. jejuni*	*Campylobacter*	*C. jejuni*	*Campylobacter*	*Campylobacter*	Unclassified	Unclassified	Unclassified
147 (*C. jejuni*)	*C. jejuni*	*C. jejuni*	*C. jejuni*	Unclassified	*Campylobacter*	*C. jejuni*	Unclassified	Unclassified	Unclassified	Unclassified

F0 = DNA extracted from fresh stool at Time 0 (at time of collection). Preservation conditions: G = Stool stored in broth with glycerol; R = Raw stool (no preservation); Z = Stool stored in Zymo DNA/RNA shield. Numbers indicate storage duration at -80°C: 1 = 1 month, 3 = 3 months, and 9 = 9 months. MDG = Metagenome-derived genome.

4.4.5 Antimicrobial resistance genotypes in metagenome-derived *Campylobacter* genomes

An AMR profile was predicted using isolates from the respective paired stool samples (Table 4.2). All MD-*Campylobacter* genomes were screened with full results tables of Table 4.5 available in Appendix 3. AMR genotype profiles in MD-*Campylobacter* genomes were frequently incomplete, especially for tetracycline resistance. Tetracycline resistance was present in 8 of the 11 sets of isolates obtained from the stool samples, while it was not identified in any of the MD-*Campylobacter* genomes at any timepoint. As observed previously, both G and R outperform Z, with minimal distinction between G and R (Table 4.5).

Table 4.5: Number of AMR Determinants Correctly Identified in Metagenome Derived *Campylobacter* Genomes Versus Isolate References for Each Stool and Storage Condition

		Preservation condition and storage time point									
Stool ID	No. of AMR determinants in Isolates	F0	G1	G3	G9	R1	R3	R9	Z1	Z3	Z9
124	4	2	1	3	1	0	1	0	0	1	1
130R1	2	0	0	0	0	0	0	0	0	0	0
130R2	2	0	0	0	0	0	0	0	0	0	0
132	4	3	3	3	3	3	3	3	3	3	3
135R1	3	1	1	1	1	1	1	1	0	1	1
135R2	3	1	1	1	1	1	1	1	1	1	1
136	1	1	1	1	1	1	1	1	0	0	0
141	3	2	0	0	0	1	0	2	0	0	0
143	2	2	2	2	2	2	2	2	0	0	0
144	5	0	0	0	0	0	0	0	0	0	0
145	12	0	0	0	0	0	0	0	0	0	0
146	1	1	1	0	0	1	0	0	0	0	0
147	3	2	1	0	1	0	1	1	0	0	0

F0 = DNA extracted from stool at collection. G = Stool stored in broth with glycerol. R = Raw stool stored as collected. Z = Stool stored in Zymo DNA/RNA shield. Numbers indicate storage duration: 1 = 1 month, 3 = 3 months, and 9 = 9 months. MDG = Metagenome derived genome.

4.4.6 Statistical analysis of storage conditions using coverage scores

Genome coverage was assessed by aligning MD-*Campylobacter* reads to isolate-derived genomes from sample-matched reference isolates. This quantitative measure served to evaluate the adequacy and comprehensiveness of metagenomic sequencing data for the purpose of characterising *Campylobacter* genotypes from MDG preserved in different conditions for different time points. Three genome coverage metrics were obtained for each sample: breadth, depth, and genome fraction. Breadth and depth were calculated by mapping Kraken extracted *Campylobacter* genus reads to a reference genome using BWA, while genome fraction was estimated from genomes assembled from Kraken extracted *Campylobacter* genus reads using QUAST. To account for variability in sequencing depth and data distribution, two normalisation approaches were applied: (1) values were normalised by 'reads in' and standardised to 10 million reads; (2) values were normalised by 'reads in', standardised to 10 million reads, and then $\log_{10}$-transformed to correct for skewed distributions. Reads-in represents the number of reads that were included in the *in-silico* human read removal pipeline, alternatively explained as the number of reads obtained from the sequencing run for a given sample. Raw data input for the statistical tests can be found in Appendix 4.

4.4.6.1 Shapiro-Wilk tests

Shapiro-Wilk tests were conducted on the datasets to assess the normality of the distribution of genome coverage scores (breadth, depth, and genome fraction) for each preservation condition (F, R, G, Z) at each time point (0, 1, 3, 9 months) and to inform the selection of further statistical tests. This was carried out for the data normalised and $\log_{10}$ transformed. The results of the Shapiro-Wilk tests indicated non-normal distributions (P <0.05) within the datasets. The $\log_{10}$ transformation approach was applied to improve the symmetry and reduce the impact of outliers. Based on the Shapiro-Wilk tests, this technique improved the symmetry for some data sets, but overall, the data remained of non-normal distribution (Table 4.6). A full breakdown of these tests can be found in Appendix 5.

Table 4.6: Overview of Shapiro-Wilk Tests for Normality of the Distribution of Genome Coverage Scores

Dataset	Total Tests	Normal (p ≥ 0.05)	Non-Normal (p < 0.05)
Normalisation	30	7	23
log_{10}	30	13	17

Using the Shapiro-Wilk test results, the paired non-parametric Wilcoxon signed-rank test was selected to compare the storage conditions at each timepoint to the state of the sample before storage. Benjamini-Hochberg (BH) correction was selected to control the false discovery rate.

4.4.6.2 Wilcoxon signed-rank tests storage timepoints versus timepoint 0

The Wilcoxon signed-rank test was used to compare coverage values prior to storage (F0) against values from samples that had been stored in the different preservation conditions (R, G, Z) at the three timepoints (1, 3, 9 months). This was to identify if any storage condition offered a significant (P <0.05) advantage in terms of the recovery of *Campylobacter* genomes, using the coverage metrics breadth, depth, and genome fraction. When using values normalised to "reads in" and reported per 10 million reads no significant difference (P <0.05) between the storage conditions was observed. There were some preservation conditions which were trending towards significance, for example, R breadth timepoint 3. G depth timepoints 3 and 9, R depth timepoints 1, 3 and 9, and Z depth timepoints 1, 3, and 9 (Table 4.7).

When the Wilcoxon test with BH correction is carried out using values normalised by "reads in" using log_{10} transformation some significant (P < 0.05) differences are observed. R breadth timepoint 3, Z breadth timepoint 1, 3, and 9, G depth timepoint 3 and 9, R depth timepoints 1, 3, and 9, and Z depth timepoint 1, 3, and 9. There are some conditions which are trending towards significance, G breadth timepoint 9, R genome fraction timepoint 3, and Z genome fraction timepoint 9 (Table 4.8).

Table 4.7: Wilcoxon Test Results for Breadth, Depth, and Genome Fraction of MD-*Campylobacter* Genomes from Stool Stored in Different Conditions from 0-9 months, Normalised to "Reads in" and Reported per 10 Million Reads

Metric (per 10M reads)	Condition	Timepoint	N	W-statistic	Raw *P* value	Adjusted *P* value
Breadth	G	1	13	37	0.588	0.635
Breadth	G	3	13	35	0.497	0.559
Breadth	G	9	13	26	0.191	0.258
Breadth	R	1	13	32	0.376	0.461
Breadth	R	3	13	13	0.021	0.080
Breadth	R	9	13	17	0.048	0.114
Breadth	Z	1	13	17	0.048	0.114
Breadth	Z	3	13	21	0.094	0.150
Breadth	Z	9	13	20	0.080	0.145
Depth	G	1	13	20	0.080	0.145
Depth	G	3	13	13	0.021	0.080
Depth	G	9	13	12	0.017	0.080
Depth	R	1	13	13	0.021	0.080
Depth	R	3	13	12	0.017	0.080
Depth	R	9	13	12	0.017	0.080
Depth	Z	1	13	14	0.027	0.080
Depth	Z	3	13	14	0.027	0.080
Depth	Z	9	13	14	0.027	0.080
Genome fraction	G	1	13	24	0.424	0.497
Genome fraction	G	3	13	21	0.286	0.368
Genome fraction	G	9	13	39	0.685	0.711
Genome fraction	R	1	13	39	1.000	1.000
Genome fraction	R	3	13	11	0.050	0.114
Genome fraction	R	9	13	14	0.091	0.150
Genome fraction	Z	1	13	15	0.110	0.164
Genome fraction	Z	3	13	18	0.182	0.258
Genome fraction	Z	9	13	12	0.062	0.129

F0 = DNA extracted from fresh stool at Time 0 (at time of collection). Preservation conditions: G = Stool stored in broth with glycerol; R = Raw stool (no preservation); Z = Stool stored in Zymo DNA/RNA shield. Numbers indicate storage duration at -80°C: 1 = 1 month, 3 = 3 months, and 9 = 9 months. N = The number of paired observations included in the test.

Table 4.8: Wilcoxon Test Results for Breadth, Depth, and Genome fraction, of MD-*Campylobacter* Genomes from Stool Stored in Different Conditions from 0-9 Months, with log10 Transformation Applied to Values

Metric (log10)	Condition	Timepoint	N	W-statistic	Raw *P* value	Adjusted *P* value
Breadth	G	1	13	34	0.455	0.472
Breadth	G	3	13	27	0.216	0.243
Breadth	G	9	13	17	0.048	0.078
Breadth	R	1	13	23	0.127	0.164
Breadth	R	3	13	8	0.006	0.031
Breadth	R	9	13	14	0.027	0.055
Breadth	Z	1	13	9	0.008	0.031
Breadth	Z	3	13	8	0.006	0.031
Breadth	Z	9	13	12	0.017	0.038
Depth	G	1	13	16	0.040	0.072
Depth	G	3	13	12	0.017	0.038
Depth	G	9	13	10	0.010	0.035
Depth	R	1	13	12	0.017	0.038
Depth	R	3	13	11	0.013	0.038
Depth	R	9	13	9	0.008	0.031
Depth	Z	1	13	3	0.001	0.016
Depth	Z	3	13	5	0.002	0.022
Depth	Z	9	13	3	0.001	0.016
Genome Fraction	G	1	11	16	0.147	0.181
Genome Fraction	G	3	9	10	0.164	0.193
Genome Fraction	G	9	10	15	0.232	0.251
Genome Fraction	R	1	10	21	0.557	0.557
Genome Fraction	R	3	10	7	0.037	0.072
Genome Fraction	R	9	9	8	0.098	0.132
Genome Fraction	Z	1	11	12	0.067	0.096
Genome Fraction	Z	3	11	11	0.054	0.081
Genome Fraction	Z	9	10	8	0.049	0.078

F0 = DNA extracted from fresh stool at Time 0 (at time of collection). Preservation conditions: G = Stool stored in broth with glycerol; R = Raw stool (no preservation); Z = Stool stored in Zymo DNA/RNA shield. Numbers indicate storage duration at -80°C: 1 = 1 month, 3 = 3 months, and 9 = 9 months. N= The number of paired observations included in the test.

Depth is the metric most affected by the storage in this experiment. The loss in depth has little effect on genome fraction in any of the preservation conditions. There is an effect on breadth seen in the $\log_{10}$ values at all timepoints in condition Z and at timepoint 3 in condition R. Overall, the results from the Wilcoxon test suggest preservation in G (Brucella broth + 17.5% glycerol) is the best condition with Z (5:1 ratio

Zymo DNA/RNA shield) being the least favourable condition to preserving *Campylobacter* genome integrity. However, the significance (P <0.05) is only observed in $\log_{10}$ transformation of the data and only trends are observed in the data when using values normalised by reads_in and reported per 10 million reads. Notably genome fraction in not affected which suggests that impact on storage on genome assembly in not significantly affected by the loss in breadth or depth over the 9-month storage time period.

4.4.6.3 Line Graphs

Guided by the statistical analyses identifying conditions and timepoints with significant differences or emerging trends, the following sections present line graphs to visualise these patterns. The plots for coverage metric per 10 million reads are much easier to interpret, but I also include the log10 normalisation for consistency. The timepoint F0 for each stool_id represents the sequencing from the stool sample before storage.

4.4.6.3.1 Genome depth

For MD-*Campylobacter* genomes depth metric in storage condition G no significant difference (P >0.05) was observed at timepoint one versus F0 when using values normalised by read_in. When using $\log_{10}$ transformation a significant difference (P >0.05) was observed at timepoint 3 and 9. The line plots show all samples except 143 in slow decline from their F0 starting point. Stool ID 132 exhibited a marked decrease from F0 to G1, followed by an increase to G3, and a subsequent decrease to G9 (Figures 4.2 & 4.3). Examination of other metrics, including the number of reads removed by *in-silico* host depletion and sequencing yield, suggests these measures were generally balanced. DNA yield for the 132 G3 timepoint is 169.50 ng/µLl versus 54.25 ng/µL at 132 G1 and 50.52 ng/µL at 132 G9. The same rise and fall pattern was observed for 132 conditions R and Z coinciding with the same rise and fall in DNA yield. In the case of Stool ID 143, host depletion seems to have failed during F0, as evidenced by an *in-silico* human read removal rate of 89.5 %, significantly higher than the 0.03-0.05 % range observed in other samples processed with the host depletion protocol. So, the positive effect seen in 143 is likely due to host depletion rather than storage the sample.

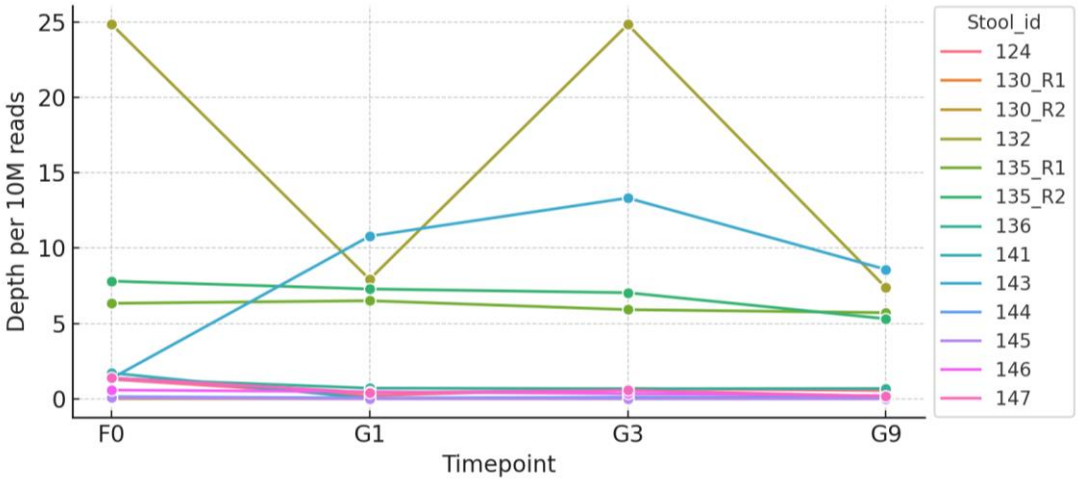

Figure 4.2: Depth per 10M reads: F0 versus G at timepoints 1, 3, and 9 months.

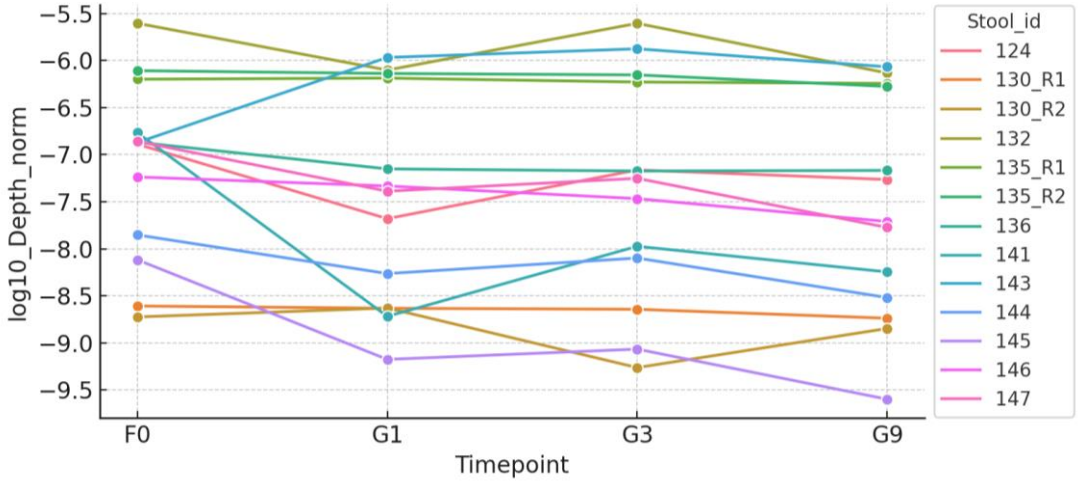

Figure 4.3: $\log_{10}$ depth normalisation: F0 versus G at timepoints 1, 3, and 9 months.

For MD-*Campylobacter* genomes depth metric in storage condition R no significant difference (P <0.05) was observed at timepoints versus F0 when using values normalised by read_in. All timepoints had a significant difference (P <0.05) versus F0 when using $\log_{10}$ values. All samples appear to be in decline from F0 to R1 in the line plots (exception 143 as previously discussed). For Stool ID 141 an increase can be seen from timepoint 3 to 9, which is much more defined in the $\log_{10}$ line graph than the normalised by reads_in version (Figure 4.4 & 4.5). The host depletion appears to have failed in the 141 R3 sample with an *in-silico* human read removal proportion at 94%, compared to R1 (64%) and R9 (42%).

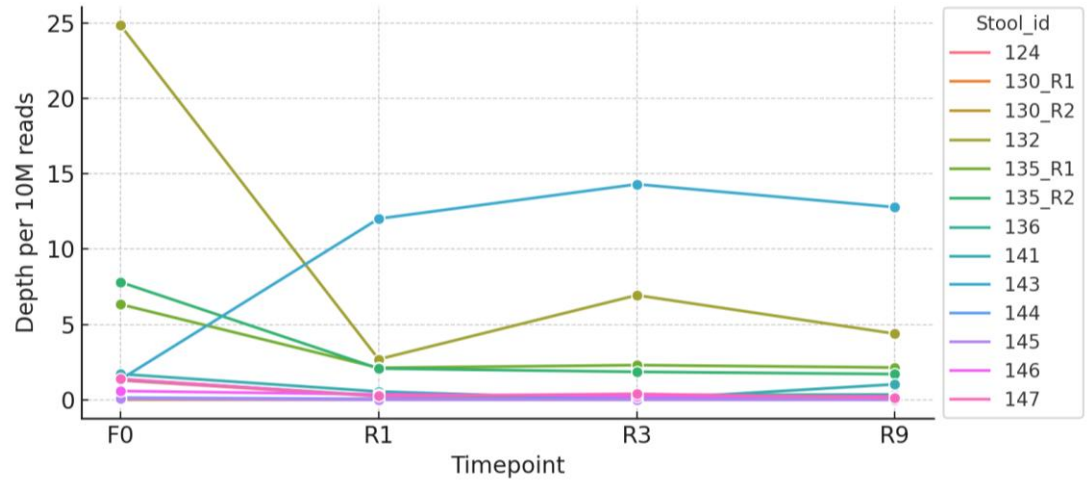

Figure 4.4: Depth per 10M reads: F0 versus R at timepoints 1, 3, and 9 months

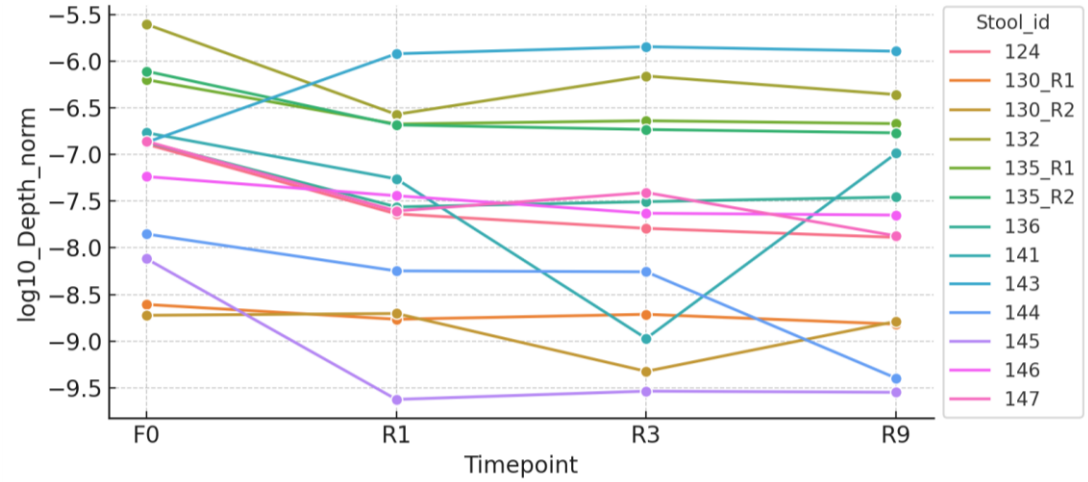

Figure 4.5: $\log_{10}$ depth normalisation: F0 versus R at timepoints 1, 3, and 9 months

For MD-*Campylobacter* genomes depth metric in storage condition Z no significant difference (P >0.05) in genome depth was observed between timepoints versus F0 when using values normalised by read_in. All timepoint did have a significant difference (P >0.05) versus F0 when using $\log_{10}$ values. An initial dip in depth can be seen in all samples apart from 132, the initial dip stabilises and remains consistent across the timepoints, this is more visually observed in the $\log_{10}$ values line graph (Figures 4.6 & 4.7). Based on the qPCR results (Section 4.3.5) and DNA extraction yields, the microbial and *Campylobacter* loads Stool ID 132 appear to be high, particularly under the Z condition.

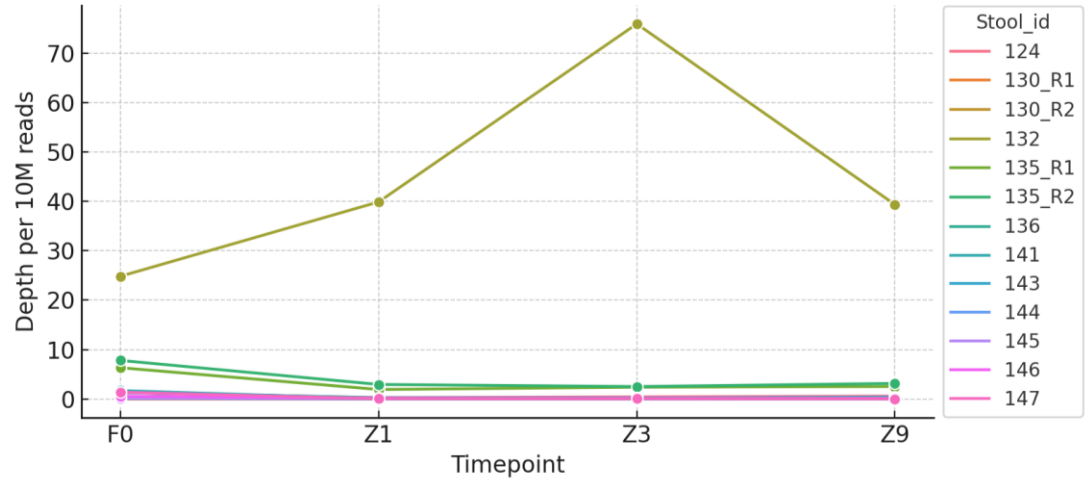

Figure 4.6: Depth per 10M reads: F0 versus Z at timepoints 1, 3, and 9 months.

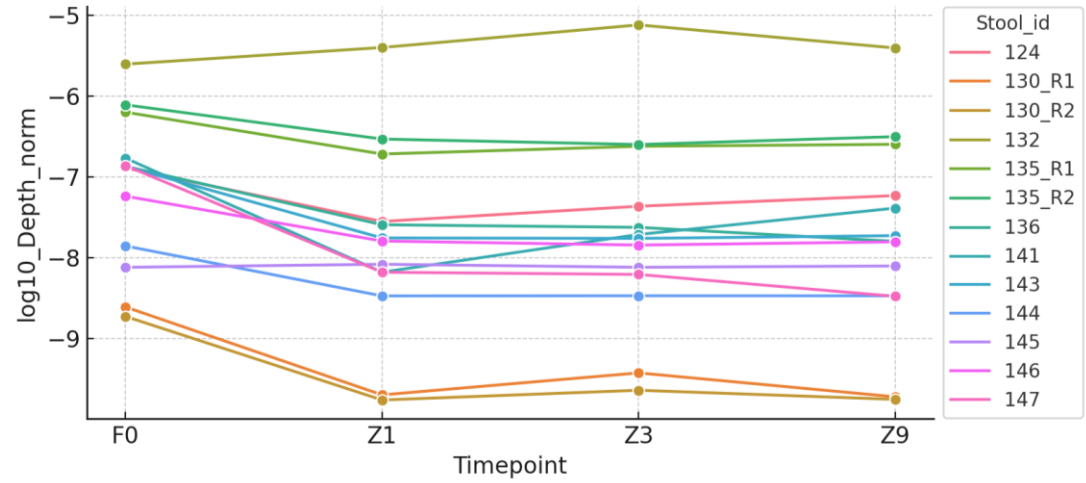

Figure 4.7: $\log_{10}$ depth normalisation: F0 versus Z at timepoints 1, 3, and 9 months.

4.4.6.3.2 Breadth

Breadth is a sequencing coverage metric that quantifies the proportion of a reference genome covered by at least one read. In contrast, depth measures how many times each base is sequenced. Breadth is particularly valuable in diagnostic applications, where detecting a larger portion of the genome can support accurate classification and typing. However, high breadth with low depth can be misleading , if each region is covered only once or very sparsely, the data may be insufficient for drawing confident conclusions.

For MD-*Campylobacter* genomes breadth metric in storage condition G no significant differences were (P >0.05) observed for breadth of the genome versus F0. For the $\log_{10}$-transformed values, timepoint 9 shows a trend towards significance (P = 0.078). Interestingly, several samples display a pattern of reduced genome breadth from F0 to R1, followed by an increase from R1 to R3 (Figures 4.8 and 4.9). While this effect does not appear as significant, there is potentially some benefit in sequencing samples were stored in condition G to analysis after a 3-month time period rather than 1 month.

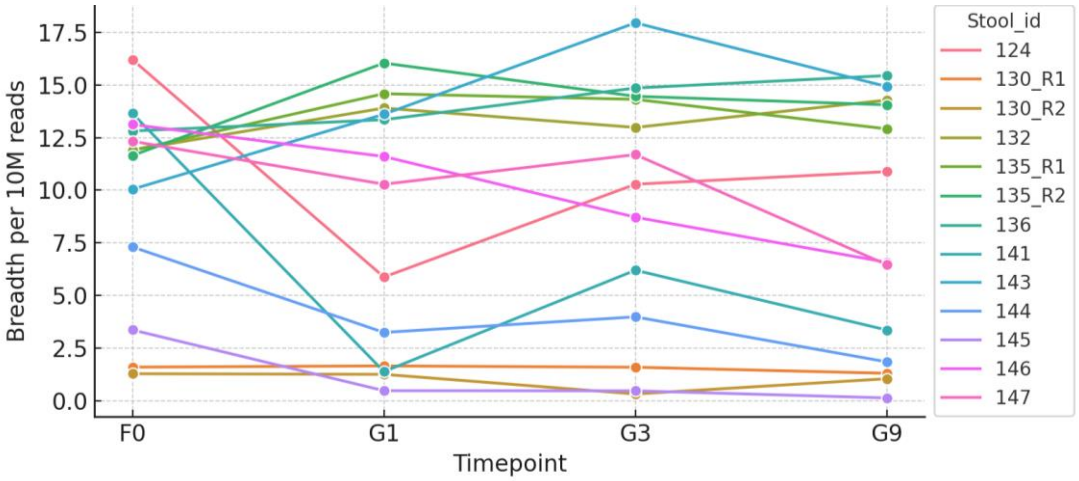

Figure 4.8: Breadth per 10M reads: F0 versus G at timepoints 1, 3, and 9 months.

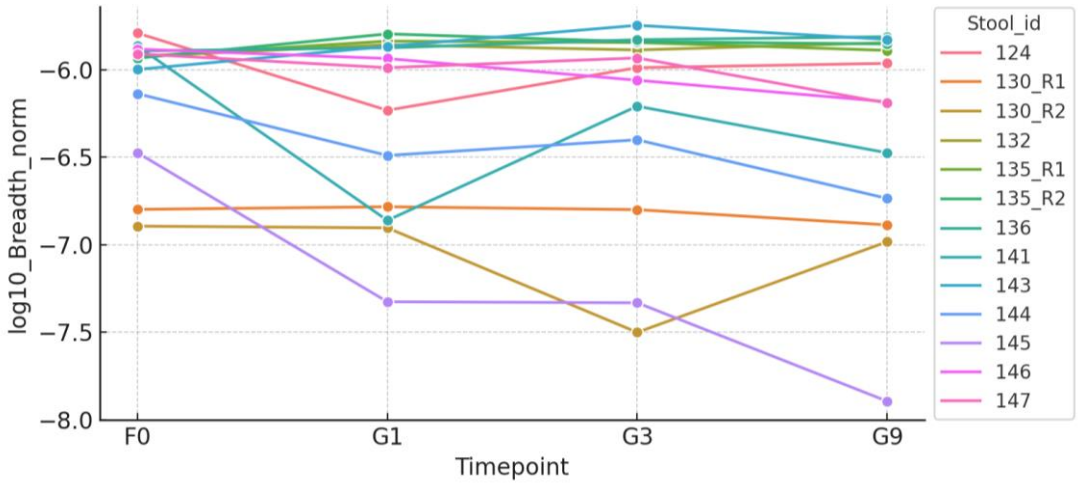

Figure 4.9: $\log_{10}$ breadth normalisation: F0 versus G at timepoints 1, 3, and 9 months.

For MD-*Campylobacter* genomes breadth metric in storage condition R no significant difference (P >0.05) in genome breadth was observed for the timepoints versus F0 when using values normalised by read_in. There is a significant difference (P >0.05) for $\log_{10}$ values F0 versus R3 and it is almost significant for F0 versus R9 (P = 0.055). Once again, the extreme bounce in stool ID 141 is observed R3 to R9 (Fig 4.10 & 4.11).

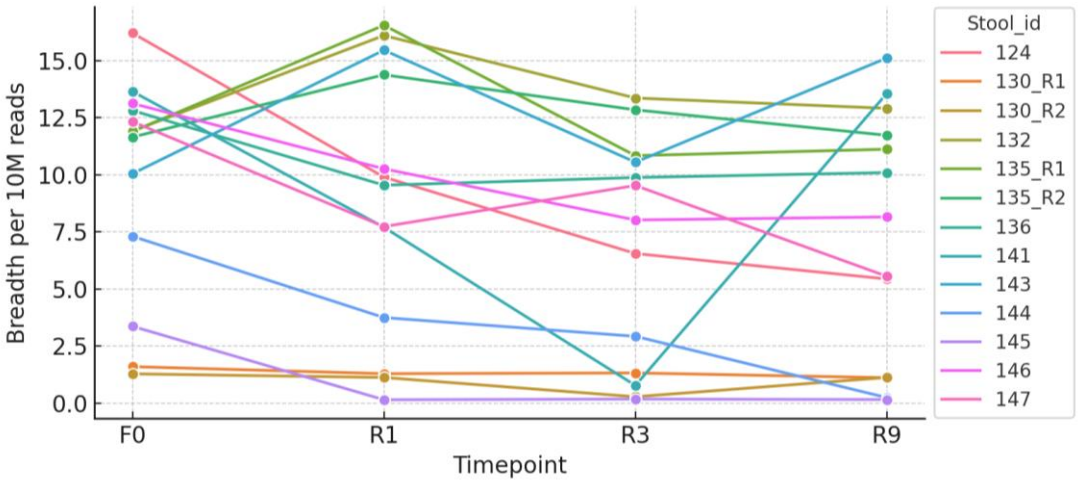

Figure 4.10: Breadth per 10M reads: F0 versus R at timepoints 1, 3, and 9 months.

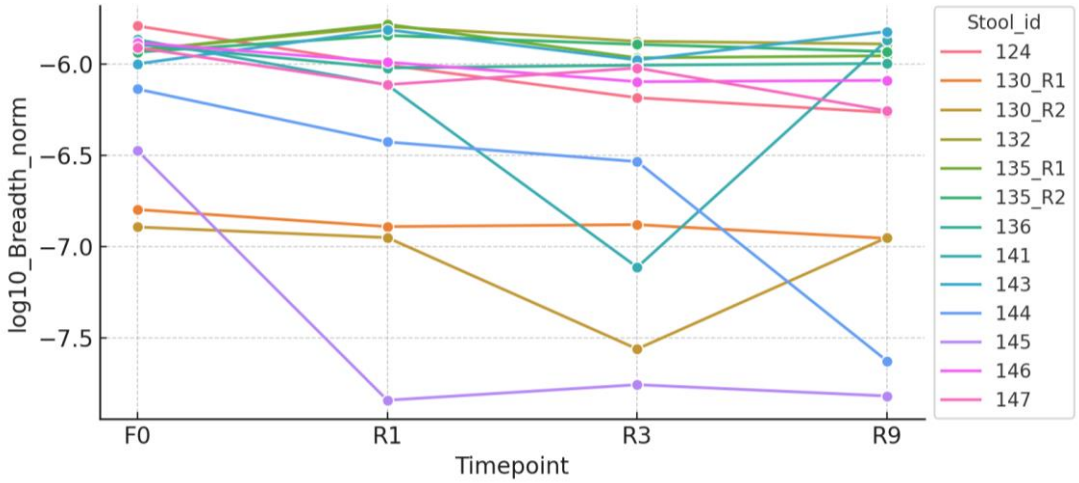

Figure 4.11: $\log_{10}$ breadth normalisation: F0 versus R at timepoints 1, 3, and 9 months.

For MD-*Campylobacter* genomes breadth metric in storage condition Z no significant differences (P >0.05) were observed between the timepoints and F0 when using values normalised by read_in. However, there was a significant difference (P >0.05) observed when comparing $\log_{10}$-transformed breadth values between F0 and timepoints 1, 3, and 9. Once the initial freeze had occurred in storage condition Z the MD-*Campylobacter* genomes retention in terms of breadth appeared relatively stable across the samples (Figs. 4.12 and 4.13).

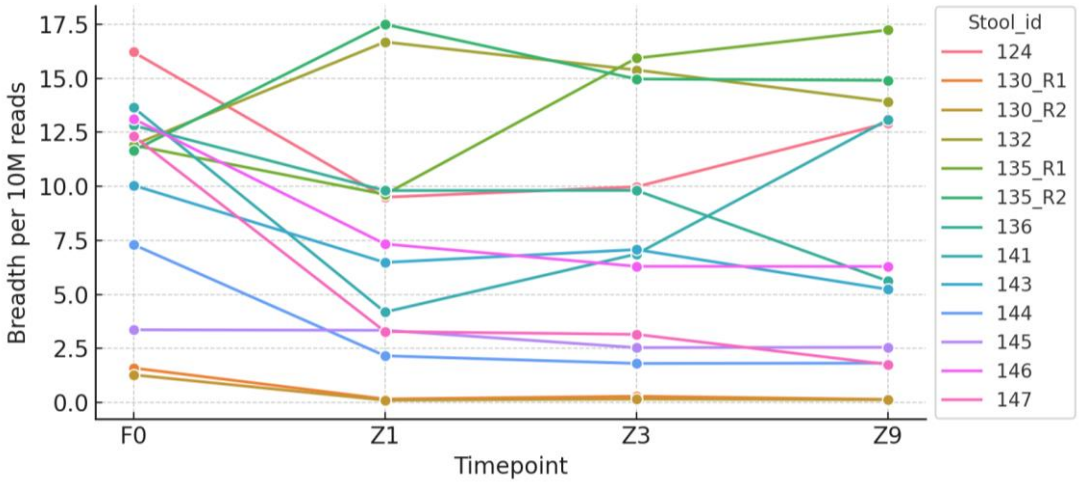

Figure 4.12: Breadth per 10M reads: F0 versus Z at timepoints 1, 3, and 9 months.

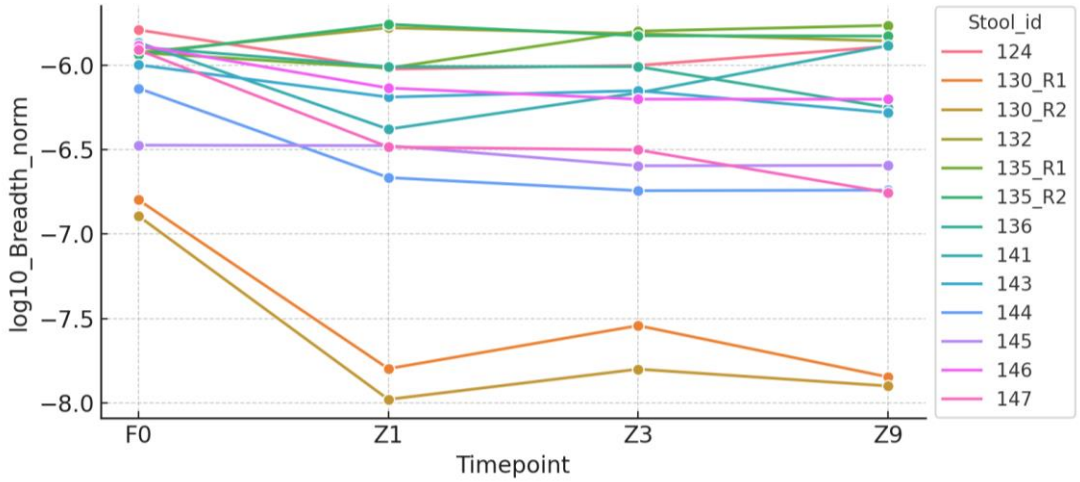

Figure 4.13: $\log_{10}$ breadth normalisation: F0 versus Z at timepoints 1, 3, and 9 months.

4.4.6.3.3 Genome Fraction

Genome fraction is a post-assembly coverage metric that reflects the proportion of the isolate derived reference genome recovered in the assembled contigs. Unlike mapping-based metrics, it excludes low-coverage or non-aligning reads that may be retained in read mapping but lost during assembly. This metric offers a more conservative estimate by minimising the influence of low-quality or false-positive reads. None of the conditions were significantly different (P >0.05) from F0 when using genome fraction values normalised by reads or $\log_{10}$ normalised values. MD-*Campylobacter* genomes genome fraction under storage condition Z exhibit a bimodal pattern: in three stool samples, preservation is maintained, while in others, a marked decline is observed between F0 and 1 month. This may relate to how Zymo DNA/RNA shield interacts with specific microbiome compositions or DNA types. MD-*Campylobacter* genomes genome fraction under storage conditions R and G preserved slightly better than those in Z. The same MD-*Campylobacter* genome fraction that preserved well in Z also preserved well in R and G, those being 124, 132, and 135. Additionally, MD-*Campylobacter* genome fraction preserved well in R and G for 136 and 143, and for G only 141. In general, the rate at which the genome fraction was lost was slower and more gradual across the 9 months in G and R. The final series of line graphs is presented in figures 4.14 through 4.19.

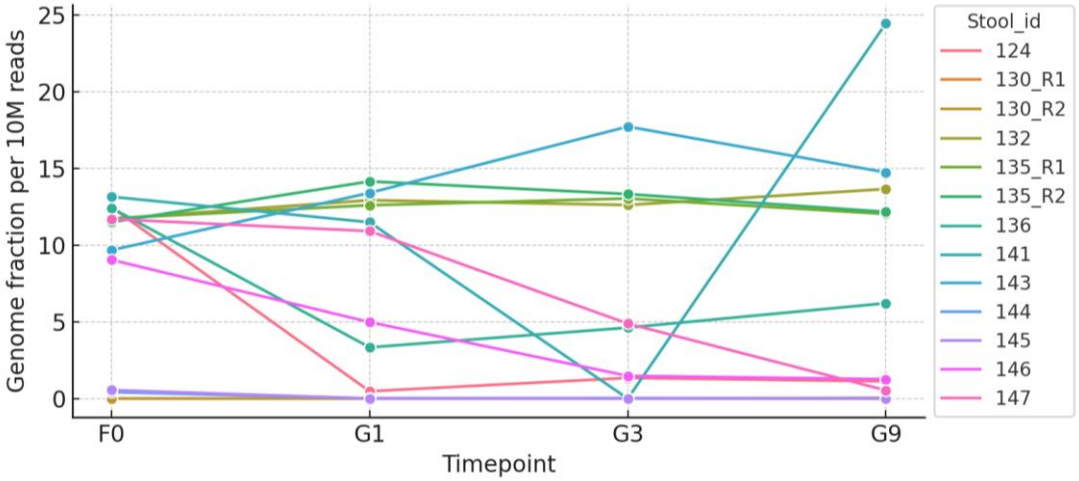

Figure 4.14: Genome fraction per 10M reads: F0 versus G at timepoints 1, 3, and 9 months.

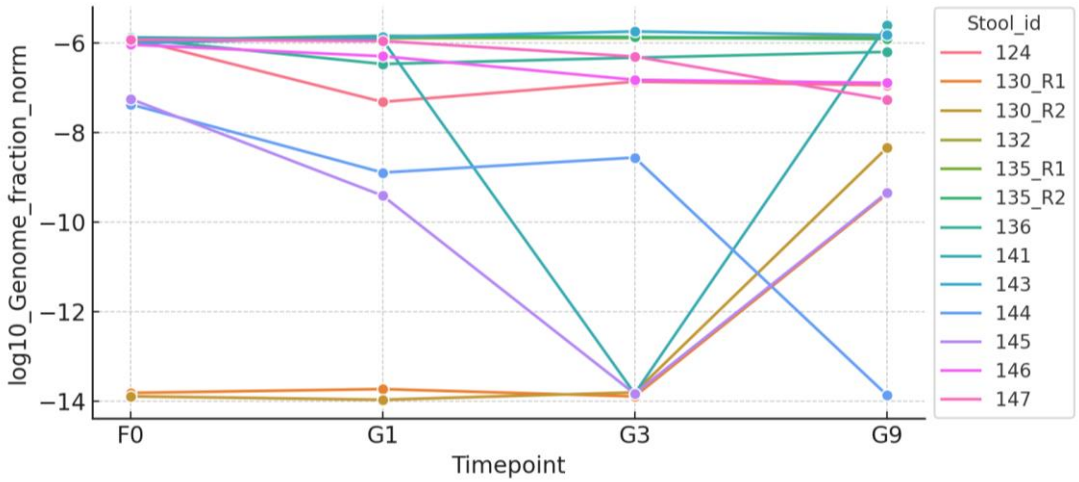

Figure 4.15: $\log_{10}$ genome fraction normalisation: F0 versus G at timepoints 1, 3, and 9 months.

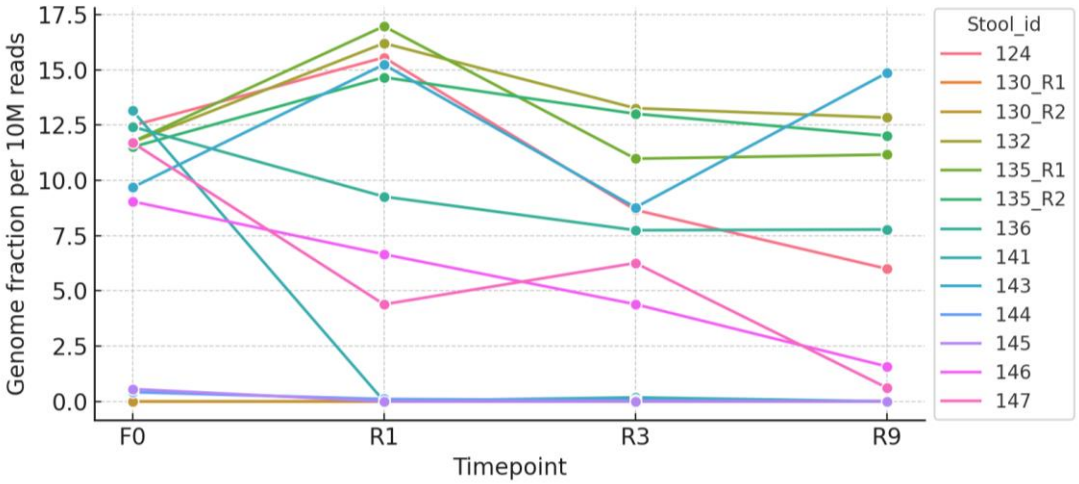

Figure 4.16: Genome fraction per 10M reads: F0 versus R at timepoints 1, 3, and 9 months.

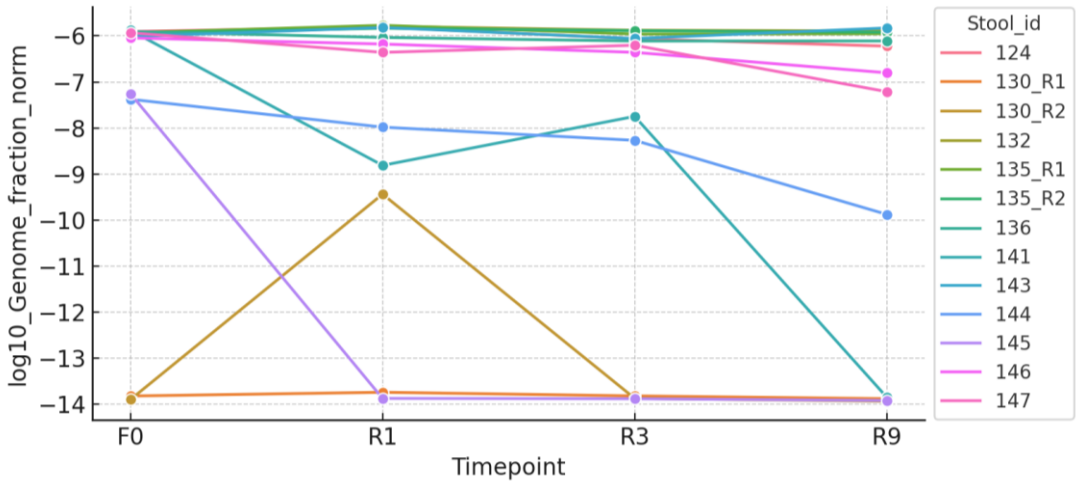

Figure 4.17: $\log_{10}$ genome fraction normalisation: F0 versus R at timepoints 1, 3, and 9 months.

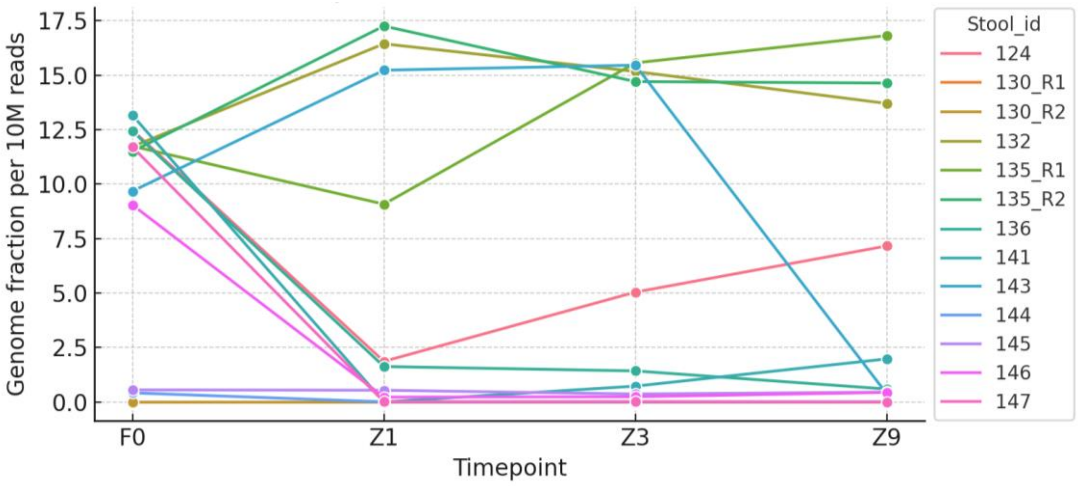

Figure 4.18: Genome fraction per 10M reads: F0 versus Z at timepoints 1, 3, and 9 months.

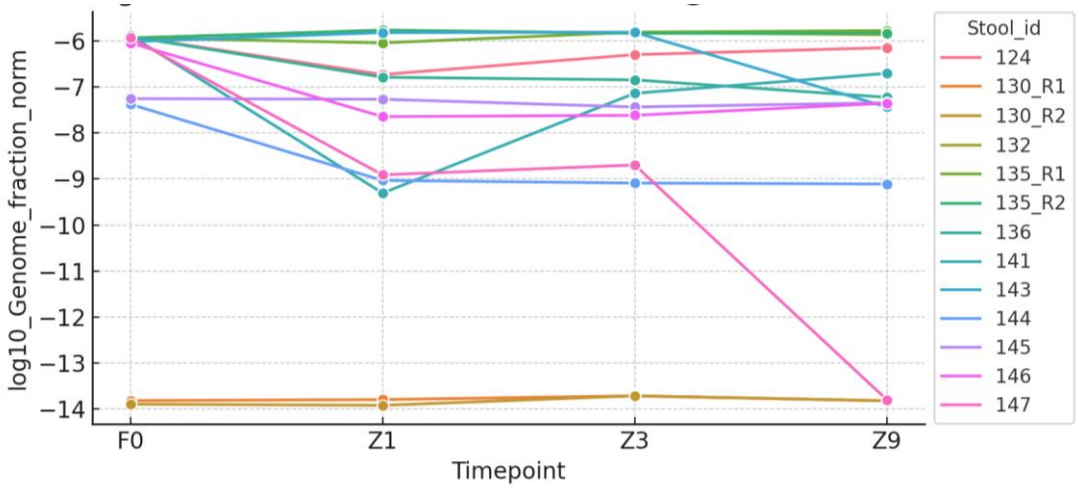

Figure 4.19: log10 genome fraction normalisation: F0 versus Z at timepoints 1, 3, and 9 months.

4.4.6.4 Conclusions: comparison of MD-*Campylobacter* genome completeness at F0 versus storage conditions over time

Sequencing depth, defined as the average read coverage across the genome, declined over time across all storage conditions when compared to the F0 baseline. Condition Z showed the most pronounced and consistent reduction, while conditions R and G exhibited similar but slightly less severe declines. Condition G retained sequencing depth more effectively in several samples, suggesting that freezing stool in a media with glycerol may offer better long-term preservation of *Campylobacter*. Despite these differences, all storage conditions showed some degree of depth loss by 9 months,

indicating a general reduction in sequencing efficiency over time. Across all storage conditions, genome breadth, defined as the proportion of the genome covered by sequencing, declined relative to the F0 baseline. This reduction was generally progressive over time. Condition Z showed the most consistent drop across replicates, suggesting that Zymo DNA/RNA Shield may limit long-term genome coverage. Conditions R and G showed more moderate losses, with some variability between samples, indicating partial preservation of genome breadth. However, none of the storage conditions fully maintained baseline levels at 9 months. All storage conditions resulted in a reduction in genome fraction recovery compared to F0, with varying degrees of severity. Conditions R and G showed modest initial declines and some stabilisation, while condition Z showed the greatest sample-to-sample variability and pronounced long-term losses in some cases.

4.4.6.5 Wilcoxon signed-rank tests assessing stability of MD-*Campylobacter* genomes once in storage

The Wilcoxon tests in the previous section were all carried out against F0 (the sample before storage). This gave an overview of the storage process, encompassing the freezing of the sample down to -80 °C. To assess the stability of the samples once stored, Wilcoxon signed-rank tests were run for timepoints 1 versus 3, 1 versus 9, and 3 versus 9. Once in storage, significant differences (P <0.05) were only observed in "Genome fraction per 10M reads" and "$\log_{10}$ Genome fraction" for MD-*Campylobacter* genomes from condition R, timepoint 1 versus 9. This significance was not observed for timepoint 1 versus 3, it became a trend for "$\log_{10}$ Genome fraction" timepoint 3 versus 9. This manifested as a reduction in genome fraction over time where the trends suggest losses between timepoints 1 vs 3 and 3 vs 9 and significance loss between timepoints 1 vs 9 (Tables 4.9, 4.10, and 4.11).

Table 4.9: Wilcoxon signed-rank tests timepoints 1 vs 9

Metric	Condition	Comparison timepoints	N	**W-statistic**	Raw ***P*** value	Adjusted ***P*** value
Depth per 10M reads	R	1 vs 9	13	44	0.946	0.946
Depth per 10M reads	G	1 vs 9	13	15	0.033	0.149
Depth per 10M reads	Z	1 vs 9	13	39	0.685	0.827
Breadth per 10M reads	R	1 vs 9	13	21	0.094	0.282
Breadth per 10M reads	G	1 vs 9	13	39	0.685	0.827
Breadth per 10M reads	Z	1 vs 9	13	37	0.588	0.827
Genome fraction per 10M reads	R	1 vs 9	13	0	0.003	0.027
Genome fraction per 10M reads	G	1 vs 9	13	40	0.735	0.827
Genome fraction per 10M reads	Z	1 vs 9	13	29	0.722	0.827
$\log_{10}$ Breadth	R	1 vs 9	13	19	0.068	0.306
$\log_{10}$ Breadth	G	1 vs 9	13	31	0.340	0.765
$\log_{10}$ Breadth	Z	1 vs 9	13	39	0.685	0.771
$\log_{10}$ Depth	R	1 vs 9	13	35	0.497	0.771
$\log_{10}$ Depth	G	1 vs 9	13	24	0.146	0.438
$\log_{10}$ Depth	Z	1 vs 9	13	38	0.635	0.771
$\log_{10}$ Genome fraction	R	1 vs 9	13	0	0.000	0.000
$\log_{10}$ Genome fraction	G	1 vs 9	13	37	0.588	0.771
$\log_{10}$ Genome fraction	Z	1 vs 9	13	43	0.893	0.893

Table 4.10: Wilcoxon signed-rank tests timepoints 1 vs 3

Metric	Condition	Comparison timepoints	N	**W-statistic**	Raw ***P*** value	Adjusted ***P*** value
Depth per 10M reads	R	1 vs 3	13	13	0.021	0.149
Depth per 10M reads	G	1 vs 3	13	36	0.542	0.885
Depth per 10M reads	Z	1 vs 3	13	44	0.946	0.946
Breadth per 10M reads	R	1 vs 3	13	40	0.735	0.885
Breadth per 10M reads	G	1 vs 3	13	36	0.542	0.885
Breadth per 10M reads	Z	1 vs 3	13	41	0.787	0.885
Genome fraction per 10M reads	R	1 vs 3	13	9	0.033	0.149
Genome fraction per 10M reads	G	1 vs 3	13	28	0.657	0.885
Genome fraction per 10M reads	Z	1 vs 3	13	27	0.594	0.885
$\log_{10}$ Breadth	R	1 vs 3	13	14	0.027	0.243
$\log_{10}$ Breadth	G	1 vs 3	13	37	0.588	0.662
$\log_{10}$ Breadth	Z	1 vs 3	13	36	0.542	0.662
$\log_{10}$ Depth	R	1 vs 3	13	45	1.000	1.000
$\log_{10}$ Depth	G	1 vs 3	13	29	0.273	0.491
$\log_{10}$ Depth	Z	1 vs 3	13	27	0.216	0.486
$\log_{10}$ Genome fraction	R	1 vs 3	13	18	0.057	0.256
$\log_{10}$ Genome fraction	G	1 vs 3	13	36	0.542	0.662
$\log_{10}$ Genome fraction	Z	1 vs 3	13	26	0.191	0.486

Table 4.11: Wilcoxon signed-rank tests timepoints 3 vs 9

Metric	Condition	Comparison	N	**W-statistic**	Raw ***P*** value	Adjusted ***P*** value
Depth per 10M reads	R	3 vs 9	13	43	0.893	0.929
Depth per 10M reads	G	3 vs 9	13	22	0.110	0.330
Depth per 10M reads	Z	3 vs 9	13	36	0.542	0.813
Breadth per 10M reads	R	3 vs 9	13	21	0.094	0.330
Breadth per 10M reads	G	3 vs 9	13	7	0.005	0.045
Breadth per 10M reads	Z	3 vs 9	13	34	0.455	0.813
Genome fraction per 10M reads	R	3 vs 9	13	15	0.203	0.457
Genome fraction per 10M reads	G	3 vs 9	13	38	0.635	0.816
Genome fraction per 10M reads	Z	3 vs 9	13	32	0.929	0.929
$\log_{10}$ Breadth	R	3 vs 9	13	39	0.685	0.771
$\log_{10}$ Breadth	G	3 vs 9	13	21	0.094	0.282
$\log_{10}$ Breadth	Z	3 vs 9	13	32	0.376	0.677
$\log_{10}$ Depth	R	3 vs 9	13	30	0.305	0.677
$\log_{10}$ Depth	G	3 vs 9	13	10	0.010	0.077
$\log_{10}$ Depth	Z	3 vs 9	13	41	0.787	0.787
$\log_{10}$ Genome fraction	R	3 vs 9	13	12	0.017	0.077
$\log_{10}$ Genome fraction	G	3 vs 9	13	39	0.685	0.771
$\log_{10}$ Genome fraction	Z	3 vs 9	13	36	0.542	0.771

4.4.7 qPCR

To complement the read coverage data, qPCR assays for *Campylobacter* (*cadF*) and human DNA were conducted. These qPCR results were paired with ST results to identify Cp thresholds indicative of when stool DNA sequencing would yield epidemiologically relevant information. An ST was obtained from isolates cultured from each stool sample (Table 4.2 & Appendix 6). Each MD-*Campylobacter* genome was screened using the same MLST tool and assigned an ST score ranging from 0 to 7. A score of 7 indicated that all seven alleles were successfully identified, allowing for assignment of a ST. Scores below 7 did not yield an ST designation but reflect the number of correctly identified alleles. Full MLST results for each MD-*Campylobacter* genome can be found in Appendix 7. A boxplot of *cadF* Cp for ST score = 7 and ST score <7 shows clear separation between the two. This clearly shows that qPCR can be used as an indicator for predicting successful sequencing typing from a MD-*Campylobacter* genome prior to metagenomic sequencing (Fig 4.20). Across all samples 42 had an ST score = 7, 88 samples had a ST score <7. For a ST Score = 7 the mean Cp was 24.99,

with a range of 19.15 to 29.10. The LOD was a Cp of 29.10. The human DNA assay also provides a useful indicator for successful sequencing typing; the separation on a boxplot is not as clear as *cadF* gene quantification, but still present (4.21). For a ST = 7, the mean Cp of human DNA was 31.78, with a range of 24.75 to 34.67. The LOD was 24.75 suggesting samples with human DNA above this threshold become troublesome for classification to the ST level.

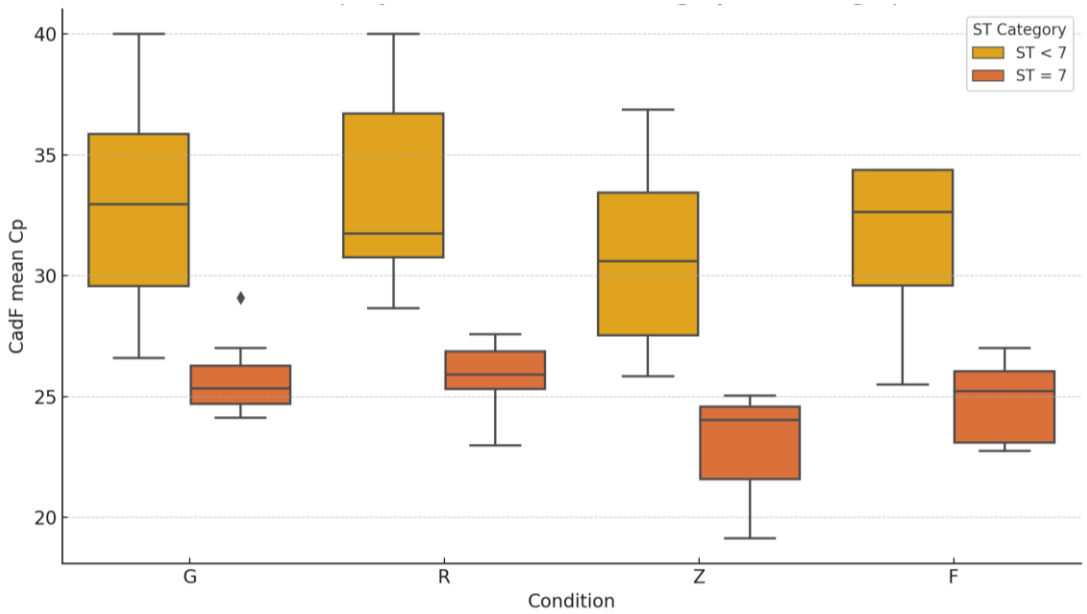

Figure 4.20: Boxplot for mean Cp *Campylobacter* (*cadF*)DNA qPCR assay separated by condition and for each condition separated by recovery of full ST score (=7) and incomplete ST score (<7), data shown for G, R, Z includes all storage timepoints (1, 3 and 9 months). F is a single timepoint, before storage (F0).

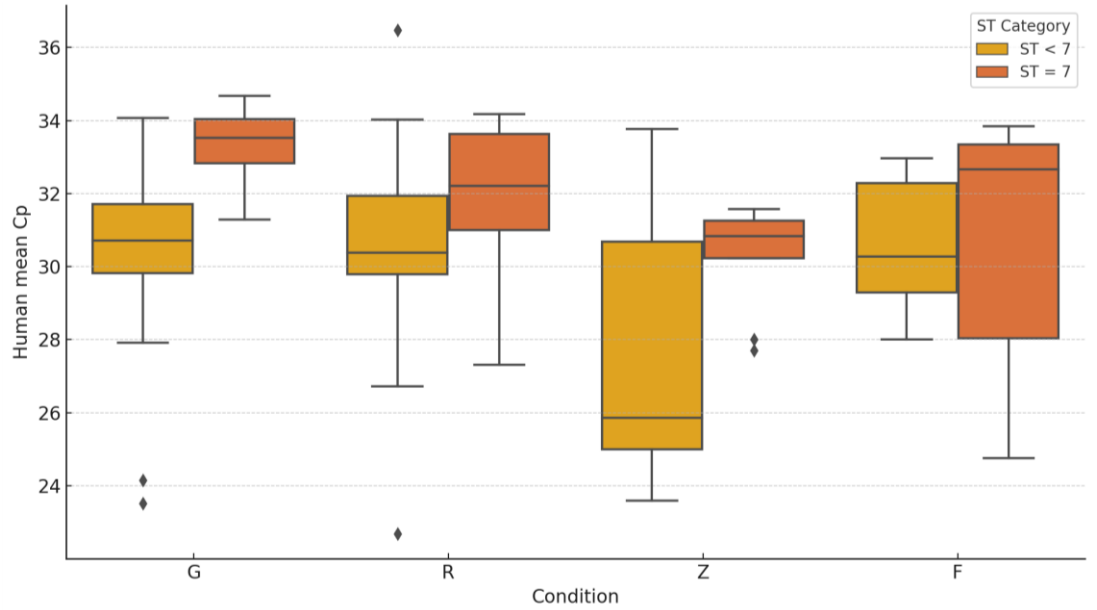

Figure 4.21: Boxplot for mean Cp Human DNA qPCR assay separated by condition and for each condition separated by recovery of full ST score (=7) and incomplete ST score (<7), data shown for G, R, Z includes all storage timepoints (1, 3 and 9 months). F is a single timepoint, before storage (F0).

Using the qPCR and ST data, a statsmodels multivariable logistic regression was carried out, inputting condition, timepoint, *cadF* mean Cp, and human mean Cp, predicting the likelihood of binary outcomes ST score = 7 or ST score <7. The conditions (G, R, Z) were input as categorical variables, allowing referencing against condition F0, *cadF,* mean Cp, human mean Cp, and timepoint were continuous variables. The result is a prediction rather than a statistical inference and gives a unit referred to as log-odds where the Odds Ratio=$e^{coefficient}$. One result from the logistical regression was significant. CadF mean cp was strongly associated with lower odds of ST Score = 7 with a p-value 0.0001. This makes perfect sense, more *Campylobacter* DNA (lower Cp) increase chances of ST score = 7.

I was concerned about the influence of the two sets of replicates in the regression model, so I repeated it using only R1 values for stool_id's 130 and 135. The CadF mean cp result remained significant (p-value 0.0001). However, removing the replicates did make another variable significant. The negative effect of storage in condition Z became significant (p-value 0.0484) (Table 4.12).

Table 4.12: Comparison of multivariable logistic regression model with and without 130 and 135 replicates

	With Reps			Without Reps		
Feature	Coefficient	Odds Ratio	p-value	Coefficient	Odds Ratio	p-value
Condition_G	–0.2989	0.7416	0.8883	–0.3496	0.7049	0.8665
Condition_R	1.6688	5.3056	0.4557	1.434	4.1953	0.5113
Condition_Z	–3.8654	0.021	0.0599	–4.1532	0.0157	0.0484
CadF mean Cp	–1.9896	0.1367	0.0001	–1.8607	0.1556	0.0001
Human mean Cp	0.3033	1.3543	0.1882	0.2798	1.3228	0.2159
Timepoint	–0.0361	0.9646	0.5706	0.0956	1.1003	0.5854

4.4.8 N50

N50 is the length of the shortest contig (or scaffold) such that 50% of the total assembled genome length is contained in contigs of this length or longer. The N50 values provide a valuable metric for assessing the quality of genome assemblies and MD-*Campylobacter* genomes. Condition G yielded the highest N50 values. However, this does not translate to advantages in classification and sequence typing. Conditions R and Z look very similar when plotted as a boxplot, suggesting Zymo DNA/RNA shield offers no protective advantage when it comes to *Campylobacter* diagnostics (Figure 4.22).

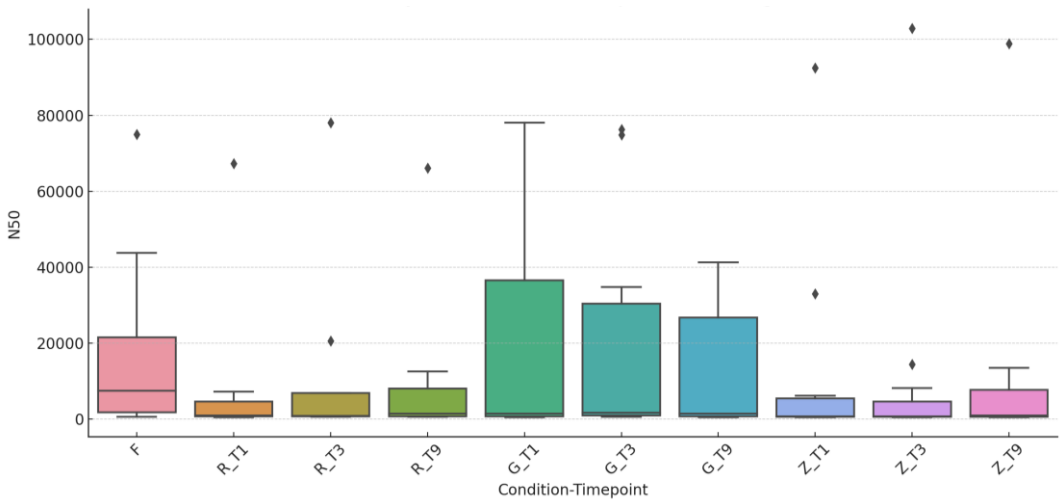

Figure 4.22: *Campylobacter* MDG N50 by preservation condition and time point.

When the data is separated by the ST score result and N50 is plotted, an N50 of 2021bp is predicted to be the limit of detection for a complete ST assignment (Fig. 4.23). This result helps to explain why the larger N50s present condition G samples did not result in better sequencing typing results compared to condition R. It is also worth noting that the mean lines for G compared to R and Z in Figure 4.22 appear similar suggesting when ample *Campylobacter* DNA is present in the stool it provides an advantage in N50 terms for storage but does not improve results when *Campylobacter* is present at low abundance.

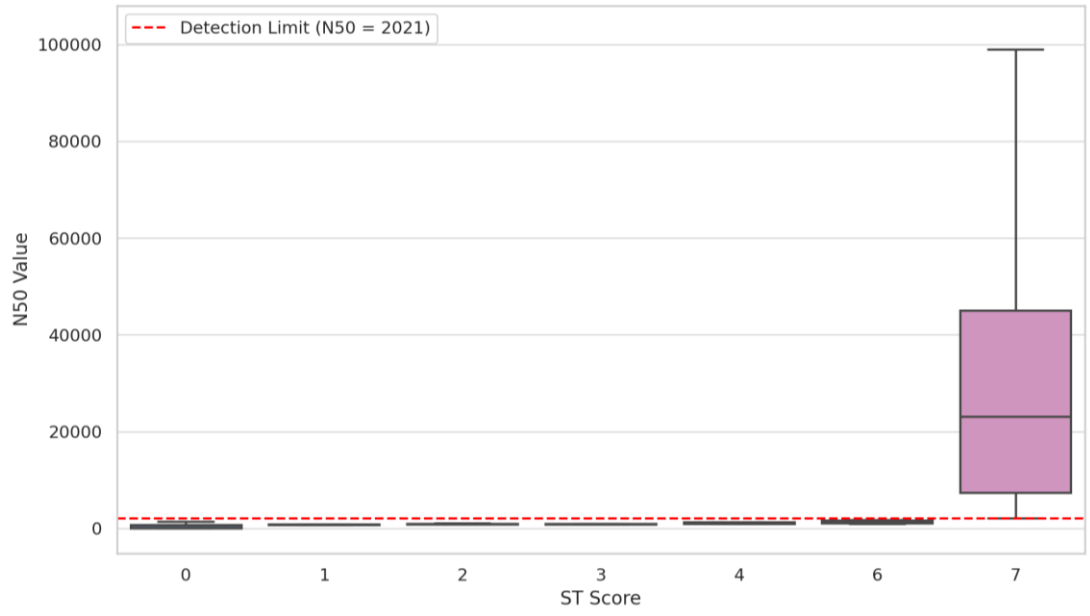

Figure 4.23: Distribution of N50 by ST score, including a limit of *Campylobacter* detection line.

4.4.9 Human Host DNA contamination

Samples stored in Zymo DNA/RNA Shield (condition Z) could not undergo human host DNA depletion, as the storage buffer lysed all cells prior to processing, therefore intact host cells containing human host DNA and free-form DNA could not be removed from the sample prior to full sample DNA extraction. Host depletion was also not performed at the time of collection, as this would not reflect a realistic workflow. In typical clinical and experimental study scenarios, samples are stored upon collection in various preservation conditions and at a storage temperature of either -20°C or -80°C and DNA extraction is performed at a later time point. There are several samples in this study that I believe the host depletion failed to work efficiently due to human error and/or the

stability of HL-SAN over time; these can be seen as outliers in the F, R, and G conditions (Fig. 4.24). HL-SAN or High Level-Salt Active Nuclease is a thermostable endonuclease used during the host-depletion protocol to degrade DNA within solution, targeting human DNA. Another potential cause of outliers was intermittent malfunction of the Eppendorf ThermoMixer, in which mixing occasionally ceased. This indicates that consistent mixing during the host depletion enzymatic step may be important for optimal performance. From the boxplot results, storage in condition R and G had a positive effect on host depletion when compared to condition F. This could help to explain the qPCR multivariable logistic regression which highlighted condition R as having a slightly improved probability of obtaining a ST score of 7 versus condition F, the pre-storage extraction.

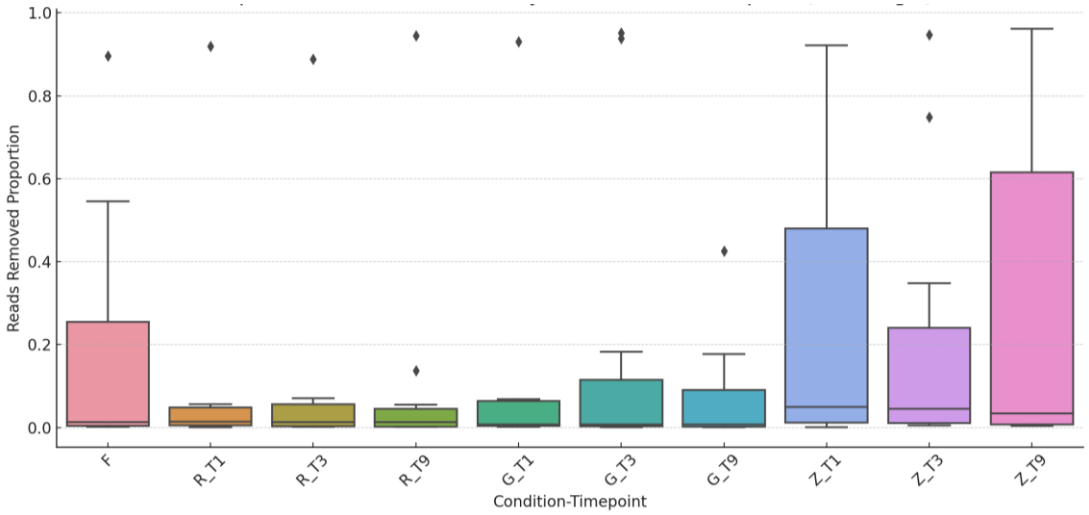

Figure 4.24: Proportion of reads removed by in-silico human read removal, a proxy for failed host depletion and human read content in the stool sample.

4.4.10 CheckM completeness

CheckM completeness forms a common metric for grading an MDG, with 85% considered high-quality, with some studies dropping this value to 70% functional analyses of microbial communities. In this study, prior to storage (F0), 7 out of 12 stool samples yielded a *Campylobacter* MDG with completeness >85%, with one MDG with completeness of ~81%, and four MDGs resulting in very poor completeness quality (<4%). To visualise CheckM completeness across time points and storage conditions, values were normalised to 'reads_in' and expressed as percentages per 10 million reads. Only stool samples exceeding 80% completeness threshold at F0 were included

in the analysis. A variety of patterns were present, with this metric displaying how challenging a stool sample can be to characterise *Campylobacter* MDGs from. For sample 124 all storage conditions negatively impact the genome completeness of MD-*Campylobacter* genomes, with R resulting in the poorest quality results. The genome completeness improved at later timepoints for conditions G and Z (Fig. 2.25).

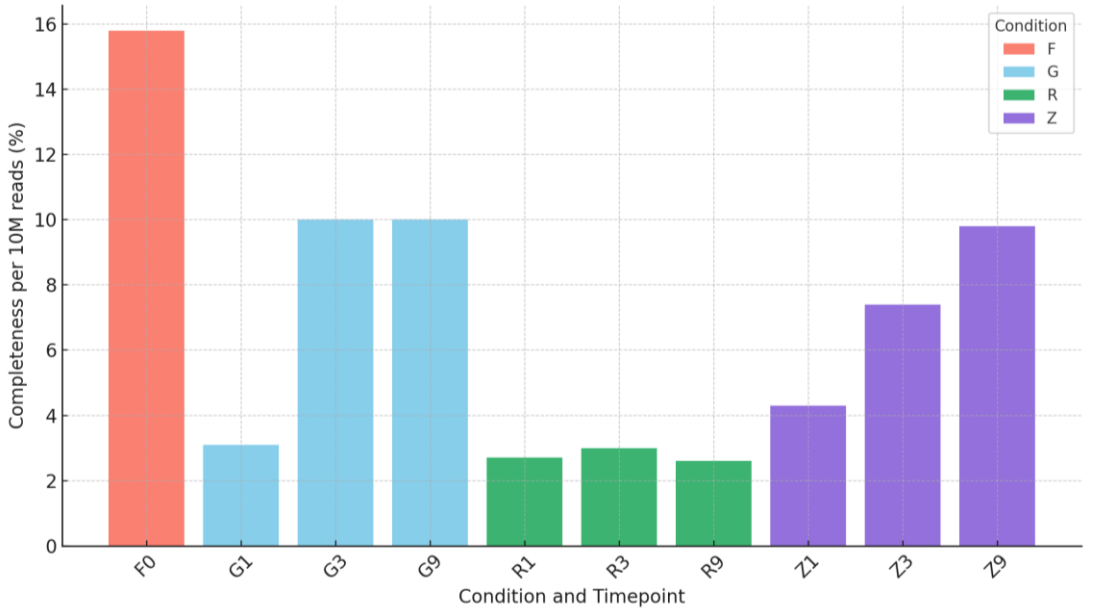

Figure 4.25: CheckM completeness of MD-*Campylobacter* genomes from stool 124. Values are percentages standardised by "reads in" and reported per 10 million reads.

For stool sample 132 all preservation conditions improved genome completeness of MD-*Campylobacter* genomes versus the pre-storage samples (F0). For R and Z the completeness was highest after 1 month in storage with a decline observed at 3 and 9 months. For G the MDG completeness fluctuated across the timepoints (Fig. 2.26). A similar pattern in present in both samples 135 replicates (135r1 and 135r2) where completeness is higher in preserved MDGs than in the pre-storage condition (F0) MDGs. Across storage conditions, distinct patterns in genome completeness were observed. In condition G, MDG completeness was higher than F0 at all timepoints in both replicates. In condition R, MDG completeness exceeded F0 after 1 month, but declined to comparable levels at 3 and 9 months. In condition Z, replicate 1 showed a progressive increase in MDG completeness across timepoints, whereas replicate 2 showed a decline (Figures 2.27 and 2.28).

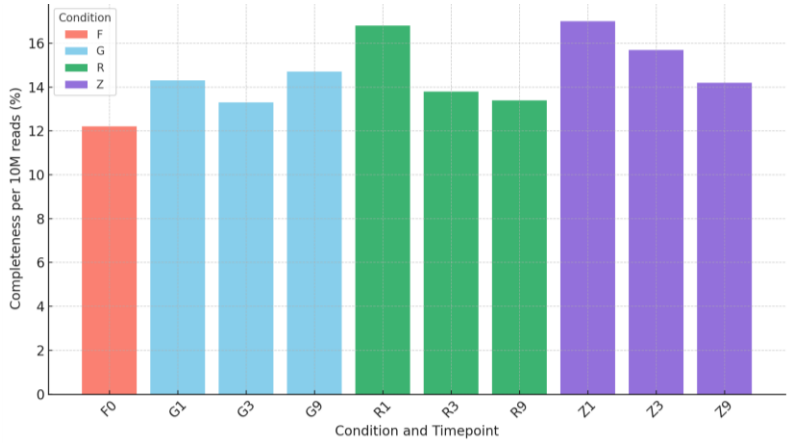

Figure 4.26: CheckM completeness of MD-*Campylobacter* genomes from stool 132. Values are percentages standardised by "reads in" and reported per 10 million reads.

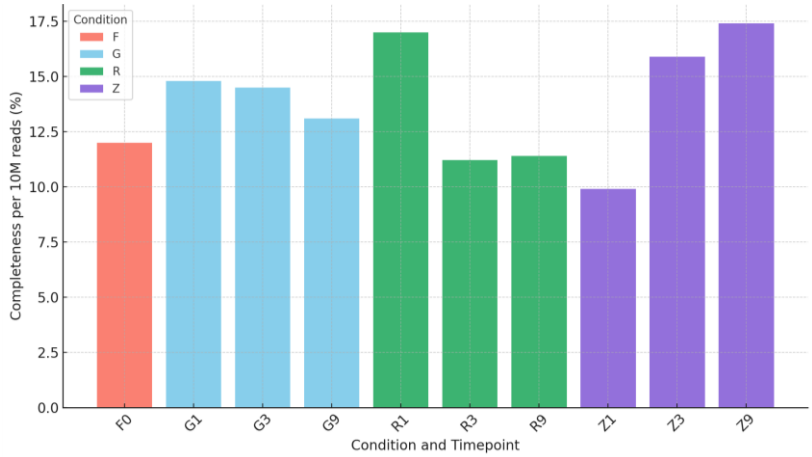

Figure 4.27: CheckM completeness of MD-*Campylobacter* genomes from stool 135 replicate 1. Values are percentages standardised by "reads in" and reported per 10 million reads.

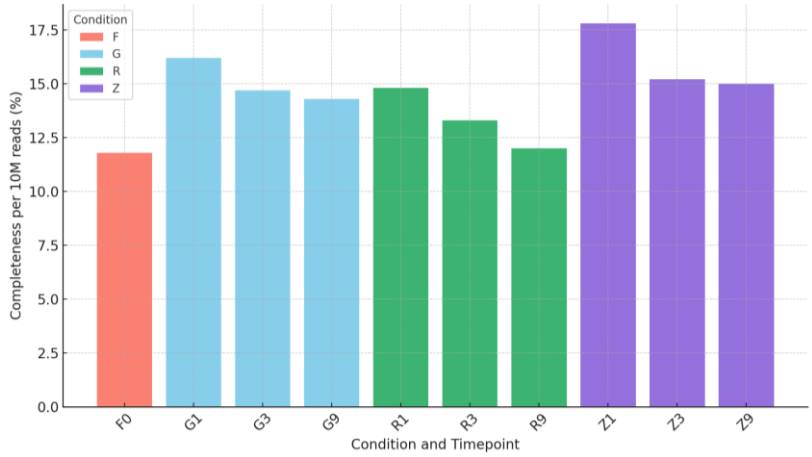

Figure 4.28: CheckM completeness of MD-*Campylobacter* genomes from stool 135 replicate 2. Values are percentages standardised by "reads in" and reported per 10 million reads.

Just as preservation at -80 °C appears to benefit MD-*Campylobacter* genome completeness, some samples start to show the opposite trend. For stool 136, recovery of MD-*Campylobacter* from G remains high; however, of MD-*Campylobacter* genome completeness for R and Z declines markedly, falling well below F0 (Fig. 2.29). For stool 141, of MD-*Campylobacter* completeness in G fails, is low in Z, and sporadic in R (Fig. 2.30).

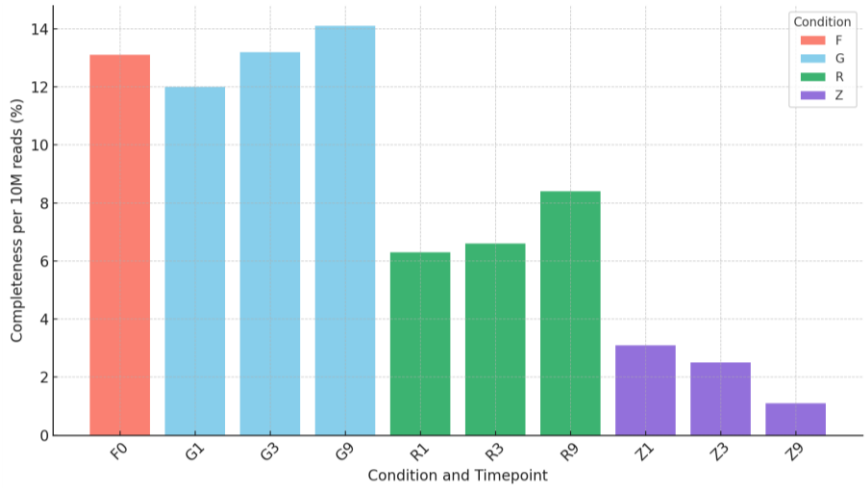

Figure 4.29: CheckM completeness of MD-*Campylobacter* genomes from stool 136. Values are percentages standardised by "reads in" and reported per 10 million reads

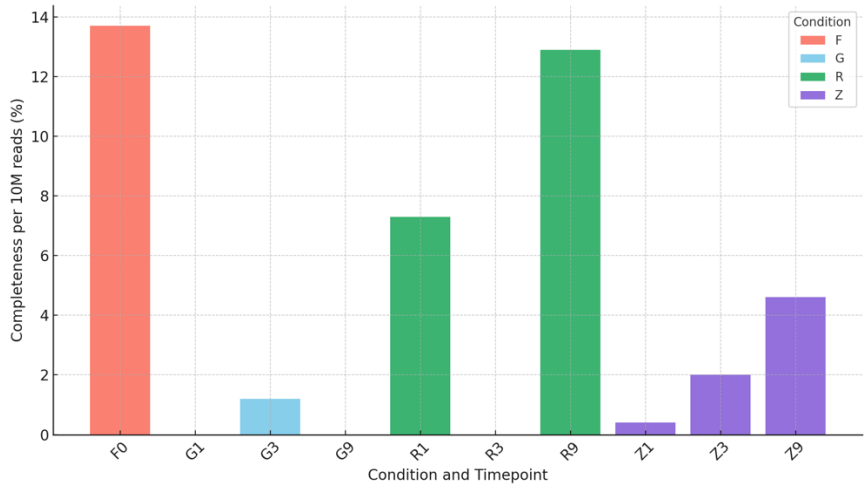

Figure 4.30: CheckM completeness of MD-*Campylobacter* genomes from stool 141. Values are percentages standardised by "reads in" and reported per 10 million reads.

In the final three stools (143, 146, and 147) the loss of MD-*Campylobacter* genome completeness recovered from Z stands out. For stool 143 MDG completeness in G and R is once again better than F0 with completeness remaining high across the time points (Fig. 2.31). For stool 146 and 147 the pre-storage (F0) MD-*Campylobacter* genomes are the most complete with Z performing poorly and G and R performing poorly by the 9-month time point (Fig. 2.32 and 2.33).

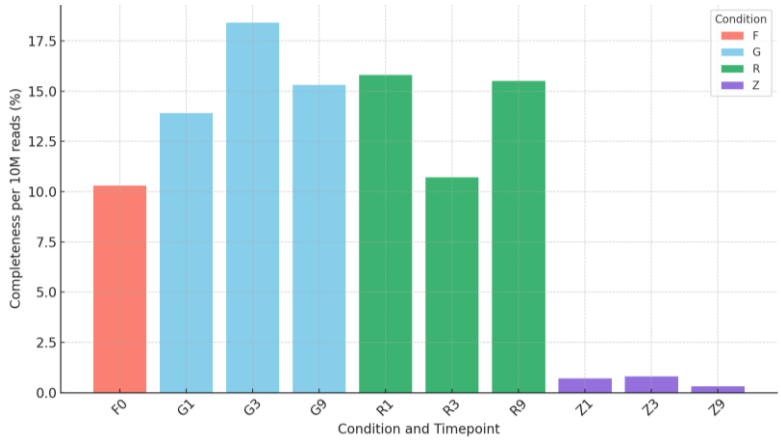

Figure 4.31: CheckM completeness of MD-*Campylobacter* genomes from stool 143. Values are percentages standardised by "reads in" and reported per 10 million reads.

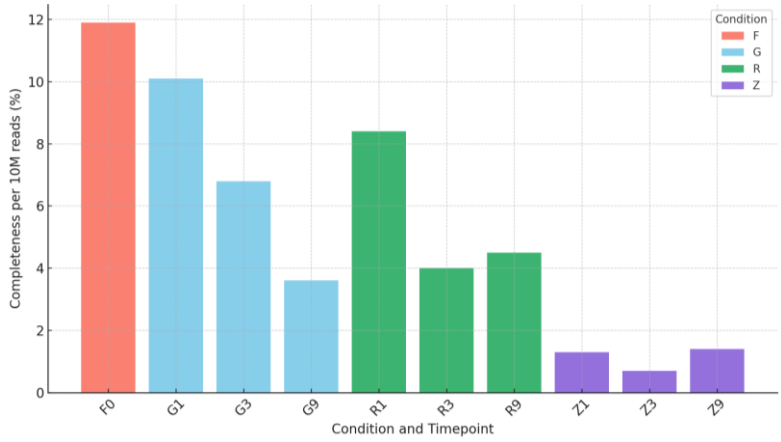

Figure 4.32: CheckM completeness of MD-*Campylobacter* genomes from stool 146. Values are percentages standardised by "reads in" and reported per 10 million reads.

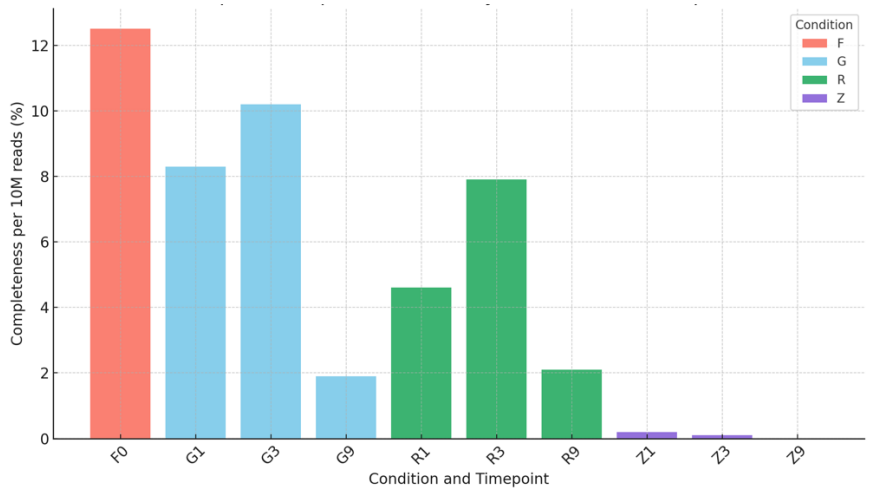

Figure 4.33: CheckM completeness of MD-*Campylobacter* genomes from stool 147. Values are percentages standardised by "reads in" and reported per 10 million reads.

In conclusion, *Campylobacter* genome completeness can be maintained, and preservation at -80°C can even be beneficial over a nine-month period when the initial the stool sample contains high levels of *Campylobacter* DNA. Predicting the *Campylobacter* DNA quality outcome in stool samples of the three tested preservation conditions was challenging, and quality metric results varied across the data set. The trend that stands out is that G and R perform as well as or better than Z in most samples. One factor to note is that the completeness score was standardised by "read_in". This shows that the *Campylobacter* DNA is present in the Z preservation condition for many samples; however, due to the high human DNA content in some samples, MD-*Campylobacter* genome coverage is lost, affecting the classification and typing metrics. As with the other coverage metric, Wilcoxon rank-sum tests (with Benjamini-Hochberg correction) were run for genome completeness, however no significant differences (<0.05) were observed for the storage conditions versus F0 (full results in appendix 8).

4.5 Discussion

4.5.1 Effect of three preservation conditions and -80°C storage for up to 9-months on *Campylobacter* detection and typing

The findings in this study demonstrate that the choice of stool preservation conditions influences the recovery of *Campylobacter* DNA for metagenomic analysis. Overall, storing stool raw (R) or in Brucella broth with 17.5% glycerol (G) was most effective at preserving sequencing utility, while Zymo DNA/RNA Shield (Z) underperformed across multiple metrics. This supports previous reports of glycerol's cryoprotective role during freezing (Gorman & Adley, 2004; Mills & Gherna, 1988), but expands on them by showing that glycerol addition does not confer a diagnostic advantage or disadvantage in the context of direct metagenomic sequencing over a 9-month time period. In most cases, R and G samples closely resembled the pre-storage baseline (F0) in genome quality and coverage, while results in Z condition often showed reduced metrics quality. The finding that DNA/RNA Shield (Z condition) significantly impairs *Campylobacter* detection and typing most notably through reduced depth and MLST recovery is strongly supported by prior research into the pitfalls of host DNA contamination (Bloomfield et al., 2023; T. Charalampous et al., 2019; Peterson et al., 2022). The current study showed that samples preserved with DNA/RNA Shield suffered from increased human DNA loads (low human Cp), reduced coverage metrics, and a drastic drop in MLST success. This aligns with Bloomfield et al. (2023), who noted that lysis-based stabilisers like DNA/RNA Shield capture excessive host DNA, overwhelming microbial signal in metagenomic data.

Coverage metrics closely mirrored these trends. Depth declined in all conditions over time, with the most significant losses observed under Z, while G showed the most consistent retention. Breadth and genome fraction were more stable, with the latter unaffected across conditions even after prolonged storage. This suggests that although samples preserved in Z condition contained lower read depth, sufficient information often remained to reconstruct the portions of the genome present in the sample. Still, only after applying $\log_{10}$ transformation did some comparisons reach statistical significance, particularly in depth and breadth measures. The downstream impact of these preservation effects was evident in genotypic outputs. Both R and G

enabled successful classification and strain-level typing in most samples. At the same time, Z frequently failed to recover enough genetic information for accurate multilocus sequence typing (MLST) or species assignment. In some instances, samples stored under R and Z conditions exhibited species misclassification, such as *C. jejuni* being erroneously identified as *C. coli*, indicating a loss of taxonomic resolution likely due to reduced sequencing quality or coverage. The lower quality metrics resulting from the Z conditions was reinforced by broader metrics of genotypic recovery, such as AMR gene detection, which similarly declined under Z storage over all time points. Taken together, these results highlight the importance of maintaining DNA integrity and minimising host contamination, both of which appear to be better achieved through raw or glycerol-based storage than through chemical buffer stabilisation when freezing samples at -80°C.

Quantitatively, coverage metrics reflected these trends. Depth of coverage was the most sensitive metric: all conditions experienced declines over time, but G showed the least loss, while Z showed the steepest declines. Breadth and overall genome fraction were less affected. Wilcoxon tests comparing storage conditions to F0 found that G tended to maintain higher normalised depth and breadth than Z, though statistical significance was generally only observed after $\log_{10}$-transform. Notably, the genome fraction (the proportion of the genome reconstructed) remained statistically unchanged across conditions, suggesting that while fewer reads mapped under Zymo, sufficient persisted to recover most of the genomic content available in the sample. In practical terms, this means that freezing with glycerol-preserved *Campylobacter* DNA yielded nearly the same results as freezing raw samples. In contrast, DNA quality in the DNA/RNA Shield preservation resulted in a significant loss of sequencing depth.

These coverage effects carried through to genotypic analyses. In multilocus sequence typing (MLST), R and G frequently recovered complete allelic profiles; for example, several samples maintained all seven loci across all time points under R and G. In contrast, Zymo-stored samples often failed to recover full MLST alleles (Table 4.3), consistent with their lower coverage. Taxonomic classification of me MDGs showed the same pattern: of 12 fresh (F0) MDGs, eight correctly identified the *Campylobacter* species, and beyond F0, there was minimal difference between R and G but both significantly outperformed Z. Some MD-*Campylobacter* genomes from the Z condition were even misclassified (e.g. as *C. coli* instead *of C. jejuni*), underscoring that storage

in DNA/RNA Shield can compromise taxonomic resolution. Logistic regression further quantified these effects: relative to the baseline (pre-storage – F0), raw-stored stool and glycerol-stored stool didn't have a significant effect on the odds of obtaining a complete MLST profile (ST score = 7), while a significant negative effect was noted for storage in Zymo. In other words, Zymo preservation markedly reduced the likelihood of complete genome typing, while raw or glycerol preservation did not.

AMR gene detection similarly suffered under Zymo storage. Overall, AMR profiles recovered from MD-*Campylobacter* genomes were incomplete, even in the best conditions. Still, R and G again allowed detection of more resistance determinants than Z. For instance, several AMR genes present in the original isolates (notably tetracycline-resistance genes) were not identified in any MD-*Campylobacter* genome, reflecting coverage gaps, chimeric structures or plasmid-located genes (Dasti et al., 2007; Hormeño et al., 2020a). Table 4.5 shows that R/G-derived MD-*Campylobacter* genomes typically contained equal or greater numbers of AMR genes compared to Z (e.g. sample 132 had 3–4 genes detected under R/G but none in Z). This pattern suggests that DNA loss during Zymo storage hampers even the recovery of small-scale genetic features. In summary, all sequence-based genome quality and typing metrics, including coverage depth, breadth, MLST alleles, taxonomic classification, and AMR loci, consistently ranked G and R as superior, with Z as the worst performer.

The findings of this study reinforce and expand upon earlier work by Loman et al. (2013), who demonstrated the feasibility of metagenomic sequencing for outbreak investigation through direct stool sequencing. Escherichia coli O104:H4 was identified and its ST was determined during a foodborne outbreak in Germany (Loman et al., 2013). Similarly, in this study, MD-*Campylobacter* genomes were used to perform MLST directly from stool. ST-level classification was achieved in samples with sufficient pathogen DNA and limited host contamination. A Key difference was Loman et al. conducted sequencing within 24 hours of collection. The exploration of preservation conditions in this chapter addresses a critical gap: whether ST-level resolution can still be achieved after prolonged frozen storage. *Campylobacter* DNA can be preserved sufficiently to yield complete ST profiles even after nine months, particularly freezing samples without a preserving agent (raw) or in glycerol-based media. This extends the utility of metagenomic ST-typing to real-world diagnostic workflows where immediate sequencing may not be possible.

Intriguingly, even though I saw higher N50 values in the G condition, this didn't lead to better ST results, nor did it lead to better AMR gene identification. This implies that while some contigs were longer, the overall genome breadth did not differ significantly between conditions, suggesting that samples with lower N50 values probably had all of the available sequence information. Similarly, Li et al. (2023) and Mehra & Kumar (2024) showed in meta-analyses that while glycerol-stored samples had improved DNA integrity metrics (e.g., fragment length), no consistent benefit was observed in microbiome or pathogen profiling resolution (Li et al., 2023; Mehra & Kumar, 2024). This confirms that while glycerol is suitable for preserving bulk DNA, its utility for direct metagenomic typing is not significantly superior to raw freezing. Moreover, the LOD for achieving a complete ST score was relatively low at 2,021 bp, meaning that even shorter assemblies could still support sequence typing. n a previous study by Djeghout et al. (2024), which explored the recovery of clinically relevant *Campylobacter* features through direct whole-genome sequencing of stool, a high proportion of samples were correctly identified to the species level when N50 values ranged from 1,000 to 10,000. Importantly, species-level identification reached 100% in samples with N50 values exceeding 10,000. The study also reported successful multilocus sequence typing in 73% of samples (n = 11), a higher proportion than observed in this study, where only 50% of F0 samples yielded a complete ST.

One striking outcome of the study is the variation in CheckM completeness across preservation conditions and samples, with G and R sometimes outperforming F0. This unpredictability is echoed by Van Zyl *et al.* (2020), who found that stool microbiota preservation outcomes are often sample-specific and not strictly condition-dependent. Harder et al. (2021) also reported long-term storage stability but highlighted that certain microbial groups (like *Campylobacter*) remain more sensitive to storage-induced degradation. These findings mirror this study, where completeness occasionally improved after storage, likely due to DNA fragment stabilisation or because the initial extraction in F0 may have been suboptimal, for example due to inefficient DNA recovery or incomplete host DNA depletion during the pre-storage processing. In most cases, degradation over time, especially in preservation condition Z, was observed. The mixed outcomes in quality metrics underlines the challenge of finding a universal stool preservation strategy. Djeghout et al. reported genus-level *Campylobacter* identification by MDG from stool was successful in 65% (24/37) of samples, versus 73% by culture and 97% by qPCR. In the 21 samples with > 60%

genome completeness, 100% were correctly assigned to species, 72% were successfully typed to STs, and 95% had accurate detection of AMR genes (B. Djeghout et al., 2024).

The qPCR data in this study provide insight into the biological factors underlying the ability to recover MD-*Campylobacter* genomes which yield clinically relevant information. Samples yielding MD-*Campylobacter* genomes with a complete ST had substantially higher *Campylobacter* DNA (lower *cadF* Cp) and lower human DNA (higher human Cp) than those that failed to yield a complete ST. In all 130 metagenomes, the mean *cadF* Cp value for successfully typed MD-*Campylobacter* genomes was 25 (range, 19–29), with a detection limit at a Cp of 29. Conversely, human DNA Cp was high (mean ~31.8) in typeable samples, with a cutoff around 24.8 (i.e. samples with more abundant human DNA, lower Cp failed to form MD-*Campylobacter* genomes). Logistic regression confirmed these effects: a higher *cadF* Cp (lower *Campylobacter* load) decreased the odds of ST recovery, while a higher human Cp (less human host DNA load) increased the odds of obtaining full ST classification. This aligns with known metagenomics principles that excess host DNA dilutes the pathogen signal (Themoula Charalampous et al., 2019; Peterson et al., 2022). Notably, the immediate cell lysis by DNA/RNA Shield (Condition Z) likely released large amounts of human DNA into the extract, which would raise the host background and suppress pathogen coverage. Indeed, this data shows that several Zymo stored stool samples had very low human Cp (high host contamination) and correspondingly poor *Campylobacter* typing results. Taken together, the results highlight host contamination as a critical confounding factor; even modest increases in human DNA can obscure low-abundance *Campylobacter* sequences, consistent with previous observations.

The results of this study align with the findings of Buytaers *et al.* (2021), who demonstrated the successful use of shotgun metagenomics to resolve a *Salmonella* Enteritidis outbreak by reconstructing pathogen genomes directly from food samples without requiring isolation. Both studies highlight the power of metagenomics to generate strain-level resolution suitable for source attribution and outbreak investigation while highlighting the detrimental effects of natural variation in pathogen load, host DNA content, and preservation effects. Buytaers *et al*. applied their workflow in an acute outbreak setting with culture-enriched food matrices, the present

study demonstrates that similar strain-level recovery is achievable from clinical stool samples stored long-term under appropriate conditions. Together, these findings reinforce the potential for culture-independent metagenomic typing to be deployed flexibly across both public health and food safety domains, even when immediate sequencing is not feasible.

In addition to its diagnostic and epidemiological applications, metagenomic sequencing has growing relevance for food safety surveillance. This is exemplified by Kocurek et al. (2023), who applied quasimetagenomic sequencing to environmental swabs from dairy and seafood production facilities and successfully reconstructed genomes from culture enrichments. Their work demonstrated that MDG from shotgun metagenomic data can achieve single-nucleotide polymorphism (SNP)-level resolution comparable to that of isolate whole-genome sequencing, enabling effective pathogen subtyping and source tracking within complex microbial communities. Such approaches not only uncover hidden diversity and persistence of foodborne pathogens in production environments but also provide a framework for integrating metagenomic tools into routine environmental monitoring. This study did not attempt SNP analysis, however considering the combination of genome coverage scores and completeness scores it is possible to predict that SNP analysis success would follow a similar pattern to the results presented, those being somewhat stool dependant and a slow loss of genome content in storage when sufficient *Campylobacter* DNA was present in the sample. Although SNP analysis was not performed in this study, the combination of genome coverage and completeness metrics suggests that SNP-level resolution would likely follow a similar pattern largely dependent on the individual stool sample and characterised by a gradual decline in genome content over time, provided sufficient *Campylobacter* DNA was present initially.

4.5.2 Implications for Diagnostic Laboratories and Resource-Limited Settings

Findings of this study carry significant implications for clinical and reference labs that could use direct from stool metagenomic approach to diagnose and characterise key clinical attributes of *Campylobacter* (and potentially other pathogens). When immediate DNA extraction and sequencing are not feasible, the optimal storage method is to freeze stool samples at –80°C (Li et al., 2023; Nel Van Zyl et al., 2020). Based on the results presented in this chapter, storing stool either without any

preserving agent (raw) or stool preserved in broth with glycerol proved to be a simple yet effective strategy for preserving *Campylobacter* DNA over a 9-month period. This approach consistently supported higher diagnostic accuracy when sequencing was performed at later timepoints. Raw frozen stool had minimal loss of coverage and the highest likelihood of yielding complete strain typing information, even matching or slightly improving upon the results of samples processed fresh. For testing labs, this means that simply storing a stool specimen in a standard cryovial at -80°C is a reliable solution for stool storage before direct sequencing. Many diagnostic workflows could benefit from this: for example, batching samples for weekly or monthly sequencing runs, or sending frozen specimens to a central facility for sequencing, can be done without significant loss of crucial pathogen data.

The study design of this chapter was purely a sequencing-based experiment, and no attempts were made to culture *Campylobacter* from the three preservation conditions at the different timepoints. However, if preserving bacterial viability for culture recovery is also a concern (for instance, to perform phenotypic antimicrobial susceptibility testing or reflexive confirmation testing of the organism), freezing stool in Brucella broth with ~15–20% glycerol is a suitable protocol to both protect viable cells and preserve pathogen DNA integrity (Li et al., 2023; Wasfy et al., 1995). Glycerol with nutrient broth stocks are commonly used to preserve isolates; here, we show that for DNA analysis, frozen glycerol stocks can preserve stool samples almost as well as freezing stool samples with no preserving agent. Over a nine-month storage time frame, stool samples stored in glycerol and broth retained *Campylobacter* DNA nearly as effectively as stool stored at -80°C without a preserving agent, with only slight additional declines in some comparison metrics. In practice, a laboratory could aliquot stool into glycerol broth vials at collection, with one advantage being that if needed, the vial can be thawed and plated to re-isolate *Campylobacter*. The results from this study indicate that doing so will not greatly compromise metagenomic sequencing ability: about 80–90% of the genome breadth was still recoverable at 9 months in glycerol (versus ~90% in raw), and the odds of successful typing were only marginally lower than raw storage. Thus, glycerol storage provides a good compromise for laboratories that value both molecular detection and the option of culture. Glycerol is inexpensive and easy to implement with basic lab supplies, making it feasible in many settings.

As a result of this study, DNA/RNA Shield preservative for stool is not recommended when metagenomic diagnosis is of microbes is the primary goal, unless there are overriding logistical needs. While the concept of an all-in-one preservation solution that inactivates pathogens and stabilises DNA at room temperature is attractive (especially for shipping or field collection far from quick transport to labs), the data in this study show a clear reduction in the detection and characterisation of *Campylobacter* when using the Zymo DNA/RNA Shield as a preservation agent for stool. The inability to perform host DNA depletion, combined with the observed faster decay in *Campylobacter* coverage, makes this approach suboptimal. Laboratories considering DNA/RNA Shield as a preserving agent for pathogen diagnostics should weigh these trade-offs. If cold-chain storage is absolutely unavailable (e.g. in remote regions or low-resource settings where even -20°C freezers are rare), using a DNA/RNA preservative might be the only way to preserve some DNA until it can be processed. In such cases, protocols must be adjusted to account for high human DNA, therefor deeper sequencing may be required to overcome the host background, and bioinformatic filtering of human reads will be essential. Even so, there is a risk that the *Campylobacter* signal could fall below detection or produce incomplete data after long delays since this pathogen is present in metagenomes in low relative abundance even at the peak of infection (Djeghout, 2024). Therefore, for best results, laboratories should prioritise the storage of stool samples at freezing temperatures (–80°C) (Wylezich et al., 2018). If samples are collected away from the testing laboratory, it may be better to refrigerate/ice-pack samples short-term and transport the samples to a facility with a suitable freezer (Newland et al., 2021), rather than immediately stabilising in Zymo DNA/RNA Shield. The study suggests that a freeze-first approach has very few disadvantages, whereas a preserve-first approach, like in Zymo DNA/RNA Shield, can lose critical information.

From a resource standpoint, the raw or glycerol storage methods are cost-effective and require minimal specialised reagents, just cryovials and freezer space. This is advantageous for routine diagnostic labs that operate under budget constraints (Yek et al., 2022). In contrast, proprietary preservation kits add per-sample cost and, as we demonstrated, may not yield a return on that investment in terms of better data. Laboratories with intermittent sequencing access (for instance, regional labs that send batches to a central sequencer) can be confident that maintaining a -80°C archive of stool specimens will allow them to perform sequencing-based testing weeks or

months later with high accuracy. Even in outbreak scenarios or prolonged case investigations, stored stool can be revisited for sequencing with reliable results if kept frozen. The passage of time had a negligible impact on success once the storage method and DNA quantity were accounted for. This is an encouraging finding, as it implies that the DNA in a raw, preserved stool sample remains of high diagnostic grade for many months. Time-related degradation is minimal if the sample is stored correctly, so labs can implement periodic sequencing without worrying that older samples will necessarily fail sequencing quality metrics, a crucial consideration for surveillance programs and research studies collecting samples over time.

4.6 Conclusion

In conclusion, I recommend that laboratories handling *Campylobacter* in stool adopt a storage protocol of freezing samples at –80°C promptly, without the addition of any preserving agents, whenever immediate sequencing is not an option. This method best preserves *Campylobacter* DNA and allows for crucial host DNA depletion steps, yielding the highest downstream sequencing quality. If concurrent culture or long-distance transport is needed, stool can be preserved in glycerol and nutrient broth and frozen, which still maintains much of the sequence quality required for detection and characterisation. In contrast, Zymo DNA/RNA Shield, despite its preserving capability, showed clear disadvantages for *Campylobacter* diagnostics when freezing stool at -80°C and should be avoided for routine use in this context. By following these guidelines, labs can ensure that metagenomic *Campylobacter* detection and typing remain as accurate as possible even after prolonged storage, ultimately improving the reliability of culture-independent diagnostics for this gastroenteric pathogen.

5 Detection of *Campylobacter* with long-read sequencing of DNA from human stool

5.1 Introduction

Campylobacter jejuni (*C. jejuni*) and *Campylobacter coli* (*C. coli*) are leading causes of bacterial gastroenteritis worldwide (Kaakoush et al., 2015). Paradoxically, these pathogens are difficult to detect by conventional means due to challenges culturing (Leblanc-Maridor et al., 2011), and in stool typically represent a small proportion of the DNA pool (Bilal Djeghout et al., 2024). As a result, important strain information (genotype, virulence factors, resistance genes) is often lost in routine diagnostics. Long-read metagenomics offers a solution. Illumina short-read platforms produce millions of high-accuracy reads, typically at reads lengths of 150bp and 300 bp (Polonis et al., 2025). While short reads work well for many genomic activities, their restricted length makes it harder to piece together whole genomes and achieve full genome level resolution (Wick et al., 2017b). In complex stool metagenomes with closely related strains and repetitive elements such as plasmids and mobile genes, short reads often yield fragmented assemblies with thousands of contigs (Bertrand et al., 2019; Olson et al., 2019). This hampers strain-level reconstruction and makes it difficult to link mobile genetic elements, such as antimicrobial resistance genes, to their host organisms. Complete pathogen chromosomes and plasmids are rarely recovered, and repetitive regions frequently disrupt assembly (G. Benoit et al., 2024; Lapidus & Korobeynikov, 2021). Moreover, low-abundance pathogens may fail to assemble or be misclassified (Lapidus & Korobeynikov, 2021). These limitations hinder comprehensive pathogen characterisation (strain typing, virulence profiling) directly from metagenomic data.

Oxford Nanopore Technologies (ONT) and PacBio are the leading long-read sequencing platforms. They can produces reads tens of kilobases in length, with ONT uniquely capable of extending into the hundreds of kilobases (C. Kim et al., 2024). ONT devices (MinION, GridION, PromethION) can yield average read lengths of several kb up to >100 kb (Espinosa et al., 2024). These lengthy fragments can span repeating sections, which reduces assembly gaps. As a result, long-read data significantly

simplifies genome reconstruction. Fewer, longer contigs are typically sufficient to cover a genome, and complex regions such as genomic islands, transposons, and rRNA operons can often be resolved due full coverage with end-to-end anchoring in flanking DNA. In practice, ONT metagenomes have enabled complete and near-complete MDGs from gut samples, revealing many more contiguous genomes than Illumina data (Gehrig et al., 2022). As a result, a higher fraction of sequences can be functional annotated. The greater continuity also supports strain-level resolution: for example, gut metagenomic assemblies recovered by long reads have yielded sufficient sequence to ST and phylogenetic placement of pathogens (B. Djeghout et al., 2024; Landman et al., 2024).

Recent studies showed that shotgun *Campylobacter* MAGs from stool could be obtained without isolation. *C. jejuni* DNA (even when only ~1–2% of reads) yielded MAGs covering >60% of the genome; these were sufficient for accurate species ID in all cases and for ST assignment in 72–95% of assemblies, as well as precise antimicrobial resistance gene profiling (B. Djeghout et al., 2024). In clinical practice, these long-read approaches have already demonstrated value: one field report used on-site ONT stool sequencing to identify *C. jejuni* as the cause of paediatric diarrhoea within a single day (Kumburu et al., 2023). Collectively, these examples illustrate how ONT metagenomics can overcome the fastidious nature and low abundance of *Campylobacter*, recovering nearly complete pathogen genomes directly from stool and enabling the same strain-level insights normally obtained only from cultured isolates.

This project aimed to evaluate the Fire Monkey high-molecular-weight (HMW) stool DNA extraction protocol developed in Chapter 2 against the Promega Maxwell RSC Faecal Microbiome DNA Kit for detection and typing of *Campylobacter*. Both long- and short-read sequencing were carried out on two stool samples, each prepared using both extraction methods. Two stool samples (n = 2) were analysed in this chapter. The sample size was constrained by financial resources and sequencing costs; however, it was sufficient for a comparative proof-of-concept evaluation of the two extraction methods using both long- and short-read sequencing.

5.2 Aims and objectives

- Successfully sequence stool-derived DNA on a MinION platform using the Fire Monkey extraction protocol developed in Chapter 2.
- Compare the performance of the Fire Monkey and Maxwell stool DNA extraction protocols for *Campylobacter* detection and sequence typing.
- Evaluate the effectiveness of long-read versus short-read metagenomic sequencing for *Campylobacter* identification and sequence typing from stool samples.

5.3 Methods

5.3.1 Sample collection

Surplus diarrhoeal stool specimens were collected from the National Health Services (NHS) Eastern Pathology Alliance (EPA) laboratory, Norwich, Norfolk, United Kingdom (UK). Stool specimens represented two separate anonymised patients with gastroenteritis symptoms who submitted specimens to the laboratory, stool 164 on 29/11/24 and stool 165 on 03/12/2024. I collected these samples on 04/12/24. DNA was extracted on stool was streaked to media on 05/12/24. *Campylobacter* spp. were initially identified in the stool specimens by the diagnostic laboratory using a rapid automated PCR-based culture-independent testing panel (Gastro Panel 2, EntericBio, Serosep, UK). Once PCR results were confirmed, a 15-20 mL aliquot of stool was placed into a sterile specimen container and transported to Quadram Institute Bioscience triple contained.

5.3.2 DNA extraction from stool

Fire Monkey DNA extraction from stool was carried out using the protocol developed in Chapter 2 and described in 2.4.3.12. Maxwell DNA extraction from stool was carried out using the protocol described in 4.3.5. To increase the amount of DNA recovered from the Maxwell protocol two cartridges were run utilising all the lysate from each

extraction. Due to the input differences, 200 mg for Maxwell and 50mg for Fire Monkey, four 50 mg Fire Monkey extractions were run and the DNA was pooled. This was to ensure the input for sequencing was the DNA from 200 mg of stool.

5.3.3 DNA extraction from isolates

Campylobacter was isolated from stool as described in 4.3.3. DNA was extracted using Fire Monkey as described in 4.3.6.

5.3.4 DNA sequencing – Metagenome

5.3.4.1 Short-read sequencing

Library preparation was carried out using Illumina DNA prep as described in 4.3.9. Sequencing was carried out externally by Azenta on an Illumina Novaseq X

5.3.4.2 Long-read sequencing

An adapted version of Nanopore's ligation sequencing DNA V14 SQK-LSK114 protocol was followed. In my experience the best sequencing runs in terms of yield and speed at which yield occurs requires high pore occupancy from the offset of sequencing. This requires more DNA than the base protocol states. Therefore, instead of 1 µg, 3 µg was used as input into the first reaction.

For the DNA library preparation, a total reaction volume of 60 µL was prepared as follows: 48 µL of DNA, 7 µL of NEBNext FFPE DNA Repair Buffer v2, 2 µL of NEBNext FFPE DNA Repair Mix, and 3 µL of Ultra II End-prep Enzyme Mix. The reaction mixture was subjected to thermal cycling conditions: initial incubation at 20°C for 5 minutes followed by incubation at 65°C for 5 minutes. Resuspended AMPure XP Beads were added to the end-prep reaction (60 µL) and mixed by tube flicking. The mixture was incubated on a Hula mixer (rotator mixer) at room temperature for 5 minutes. Subsequently, the sample was centrifuged to pellet the beads on a magnet until the supernatant became clear and colourless. The tube remained on the magnet while the supernatant was pipetted off. The beads were then washed twice with 200 µL of freshly prepared 80% ethanol without disturbing the pellet. After each wash, the ethanol was removed using a pipette. The tube was placed back on the magnet between washes,

and any residual ethanol was pipetted off after the final wash. The beads were allowed to air dry for approximately 30 seconds. After removing the tube from the magnetic rack, the pellet was resuspended in 61 µL of nuclease-free water and incubated at room temperature for 2 minutes. The beads were pelleted on a magnet again, for at least 1 minute. Finally, 61 µL of the eluate was removed and retained into a clean 1.5 mL Eppendorf tube.

In the DNA ligation process, a total reaction volume of 100 µL was prepared as follows: 60 µL of DNA sample from the previous step was combined with 5 µL of Ligation Adapter, 25 µL of Ligation Buffer, and 10 µL of Quick T4 DNA Ligase. The reaction was thoroughly mixed by gentle pipetting and briefly spun down. It was then incubated for 10 minutes at room temperature. AMPure XP Beads were resuspended by vortexing. Subsequently, 40 µL of resuspended AMPure XP Beads were added to the reaction and mixed by flicking the tube. The mixture was incubated on a Hula mixer for 5 minutes at room temperature. The tube remained on the magnet, and the supernatant was pipetted off. The beads were washed by adding 250 µL of Long Fragment Buffer. The beads were flicked to resuspend, spun down, and then returned to the magnetic rack to allow the beads to pellet. The supernatant was removed using a pipette and discarded. This washing step was repeated once more. After the final wash, the tube was spun down and placed back on the magnet. Any residual supernatant was pipetted off, and the beads were allowed to air dry for approximately 30 seconds. Following drying, the tube was removed from the magnetic rack, and the pellet was resuspended in 15 µL of Elution Buffer. The mixture was spun down and incubated for 10 minutes at 37°C. The beads were pelleted on a magnet again for 1 minute. Finally, 15 µL of the eluate containing the DNA library was removed and retained into a clean 1.5 mL Eppendorf DNA tube.

Between 223 ng and 756 ng was loaded onto a MinION R10.4.1 flowcell. The sequencing reaction was set up with a total volume of 75 µL. This included 37.5 µL of Sequencing Buffer, 25.5 µL of Library Beads prepared immediately before use and 12 µl of the DNA library. The flow cell was flushed using a total volume of 1,205 µL. This included 1,170 µL of Flow Cell Flush, 5 µL of Bovine Serum Albumin at a concentration of 50 mg/mL, and 30 µL of Flow Cell Tether. Loading of the device was as standard.

5.3.5 DNA sequencing – Isolates

5.3.5.1 Short-read sequencing

Library preparation was carried out using Illumina DNA prep as described in 4.3.9. Sequencing was carried out by QIB sequencing on an Illumina Nextseq 500.

5.3.5.2 Long-read sequencing

5.3.6 Long-read metagenome assembly pipeline

The raw fastq files were filtered using Nanofilt (Galaxy v0.1.0) set to filter out reads less than 500 bp and/or below a Q score of 10 (De Coster et al., 2018). Assembly was carried out using metaMDBG (G. Benoit et al., 2024). The output contigs fasta from metaMDBG was binned using SemiBin2 (Galaxy v2.0.2;(Pan et al., 2022)). The cached database was gtdb_v95. The ORF finder used to estimate the number of bins was Prodigal (Hyatt et al., 2010). The human gut environment built-in model was activated. SemiBin2 required a bam file of reads mapped to the contigs, this file was produced using minimap2 (Galaxy v2.28) using the map-ont setting for mapping ONT reads (Li, 2018).

5.3.7 Short-read metagenome assembly pipeline

The MetaWrap2 (Galaxy v1.3.0) pipeline was used to create MAGs from the short-read data sets (Uritskiy et al., 2018). Megahit (Galaxy v1.2.9) was used to create an assembly input for MetaWrap2 (Li et al., 2015). Paired reads files and the assembly were used as input.

5.3.8 *Campylobacter* bin identification

CheckM (Galaxy v1.20) was used to get an identification for bins (Parks et al., 2015). Bins identified as *Campylobacter* by CheckM were extracted and the individual bins were checked with GTDB-tk (Galaxy v2.2.2;(Chaumeil et al., 2019)).

5.3.9 Kraken read recovery MAGs

An alternative strategy for MAG construction was to classify reads and use the classification to isolate the reads of the species of interest. Short-read files were processed with Fastp (Galaxy v0.23.2;(Chen et al., 2018)) and then classified using Kraken2 (Galaxy v2.1.3;(Wood et al., 2019)). Krakentools (Galaxy v1.2) was used to extract *Campylobacter* reads using Taxonomic ID 194 (Lu et al., 2022). Reads were assembled using Megahit. The same process was followed with the long read datasets, with the extracted reads assembled using Flye (Galaxy v2.9;(Lin et al., 2016)).

5.3.10 Isolate assembly

Long-read assemblies were made using Flye (Galaxy v2.9). ONT reads were first filter with Filtlong (Galaxy v0.2.0) with min length set to 1 kb and min mean quality set to 50. The long-read assemblies were polished with short reads using Polypolish (Galaxy v0.5.0) to produce hybrid assemblies (Wick & Holt, 2022).

5.3.11 Typing and Antimicrobial resistance determinant identification

Classification of isolate and MAGs was performed using GTDB-Tk. Sequence typing was conducted using MLST (Galaxy v2.16.1) with the built in *Campylobacter* scheme selected. AbriTAMR (Galaxy v1.0.14) was used for Antimicrobial resistance (AMR) detection set to detect point mutations for *Campylobacter* (Horan et al., 2022).

5.3.12 Single Nucleotide Polymorphism analysis

Snippy4 (Galaxy v4.4.3) was used for Single Nucleotide Polymorphism (SNP) analysis of isolates and MAGs (Seemann, 2015). To select a hybrid reference per stool sample for the SNP analysis an initial SNP analysis was carried out using a random isolate from the set. This was done because the hybrids were all high quality. The isolate with the lowest unaligned genome content in the SNP test analysis were selected to be the reference for the final analysis.

5.3.13 Relative abundance

Kraken2 was used to perform taxonomic classification of both long- and short-read metagenomic datasets. Subsequent manipulation of Kraken2 reports and visualisation of taxonomic profiles was carried out in Python using the pandas (v2.2.0), matplotlib (v3.10.1), and seaborn libraries (v 0.13.2). Taxonomic profiles were filtered to retain only genus-level classifications and reads assigned to the genus *Homo* were excluded. In stool sample 165, where a high proportion of human DNA was present, relative abundances were recalculated following the removal of human reads (*Homo*) to provide a more accurate representation of the microbial community.

5.3.14 Read mapping

Refer to Section 4.3.18, where the same strategy was applied.

5.4 Results

5.4.1 Sample information

Two *Campylobacter* positive stool samples, stool 164 and stool 165 were collected from the EPA laboratory. *Campylobacter* was isolated and six colonies were sequenced per stool sample. DNA was extracted from the stool samples the day after collection having been stored at 4°C overnight. Throughout this chapter I use 164fm to represent the Fire Monkey extraction of DNA from stool 164, and 165fm for stool 165, likewise the abbreviations for the Maxwell extractions are 164max and 165max.

5.4.2 Stool DNA MinION runs

A single stool sample DNA extraction was run per MinION flowcell with 8-10 Gb as the target. I was very close to achieving pure DNA for all samples based on Nanodrop 260/280 and 260/230 ratios. For 260/280 values within the range 1.8-2.0 are considered pure, and 260/230 values in the range 2.0-2.2 are considered pure. Samples 164fm and 165max could be considered pure by Nanodrop analysis. Sample 164max fell out of range for both ratios with a 260/280 of 1.76 and a 260/230 of 1.72, Sample 165fm was out of range for the 260/230 ratio with a value of 2.43 (Table 5.1).

Table 5.1: DNA Quantifications and Nanodrop Ratios Assessing Purity of Nucleic Acids

Sample	Qubit (ng/μL)	Nanodrop (ng/μL)	260/280	260/230
164fm	143.0	131.0	1.91	2.05
164max	384.0	452.0	1.76	1.72
165fm	68.4	72.4	1.90	2.43
165max	408.0	415.0	1.87	2.03

The 164fm sample was the first to be run and ~750 ng was loaded onto the flowcell. A retrospective calculation using read N50 and ng estimates 329 fmol was loaded onto the flowcell. The 164max sample suffered high DNA loss during the long fragment buffer washes, this buffer size selects for fragments above 3 kb. This lose meant 223 ng (110 fmol) were loaded onto the flowcell. For sample 165fm 600 ng (984 fmol) was loaded and for 165max 750 ng (773 fmol) was loaded (Table 5.2). The 164fm sample produced the only library to run into the 8-10 Gb range within 24 hours. This run was stopped and the flowcell was washed and re-used for the isolate sequencing in this chapter.

Table 5.2: Final Nanopore Library for Loading on the MinION

Sample	Qubit (ng)	fmol
164fm	756	329
164max	223	110
165fm	600	984
165max	750	773

TapeStation traces of the final libraries show cleaner peaks for samples 164fm and 164max, with the loss of DNA suffered by 164max during the library prep being apparent in the size of the peak (Figures 5.1-5.2). The traces for 165fm and 165max exhibited a shouldering effect, indicating greater fragmentation of the DNA (Figures 5.3-5.4). The traces underscore that while the TapeStation may not accurately size HMW DNA, it remains valuable for quality control purposes, a clean peak being most desirable (Figures 5.1-5.4).

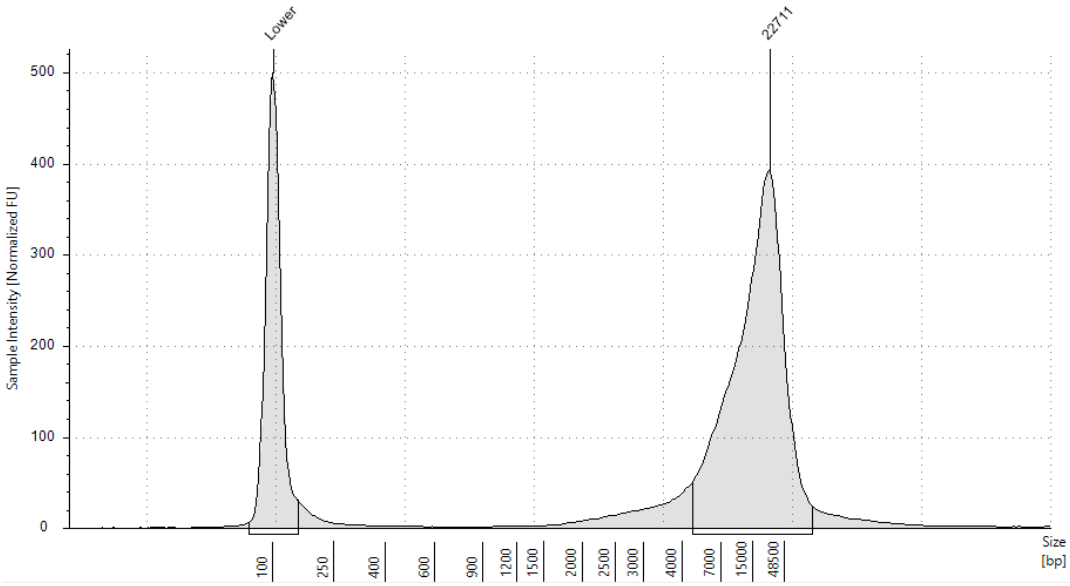

Figure 5.1: GenomeTape TapeStation trace of the Nanopore library for 164fm.

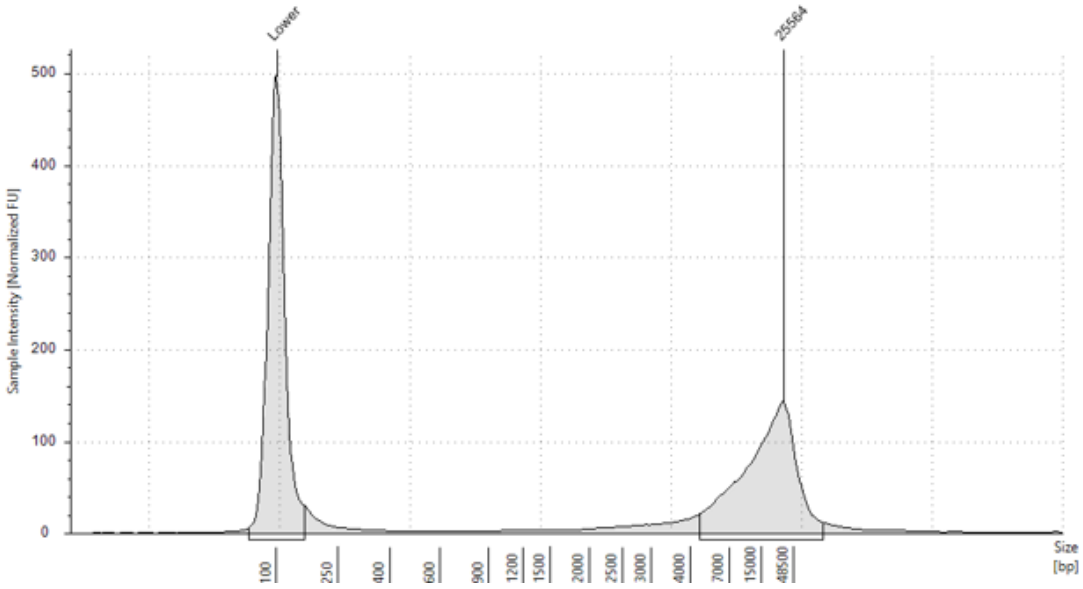

Figure 5.2: GenomeTape TapeStation trace of the Nanopore library for 164max.

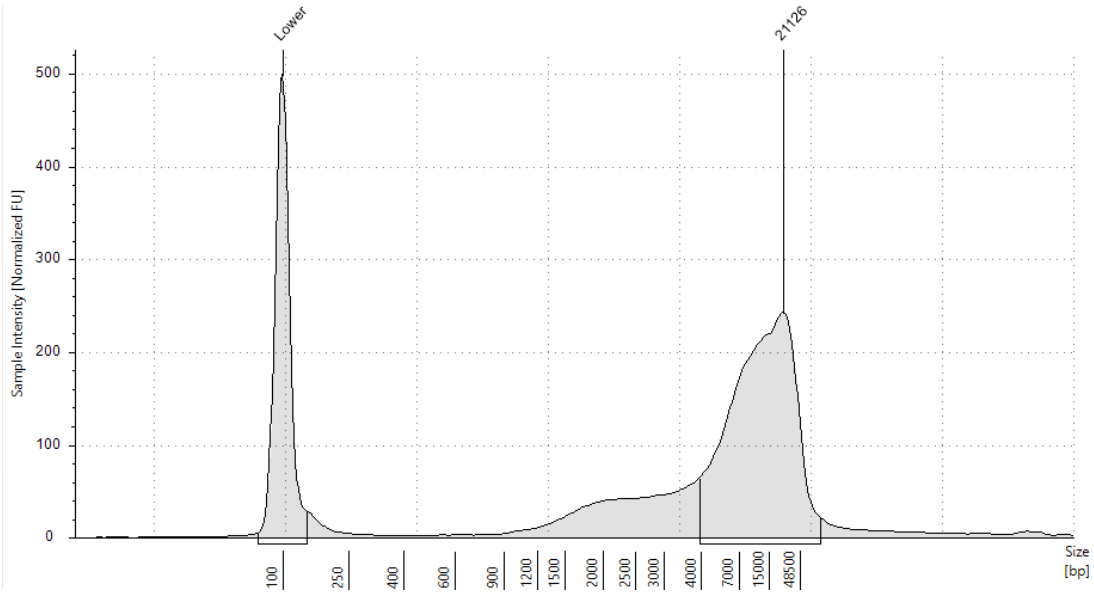

Figure 5.3: GenomeTape TapeStation trace of the Nanopore library for 165fm.

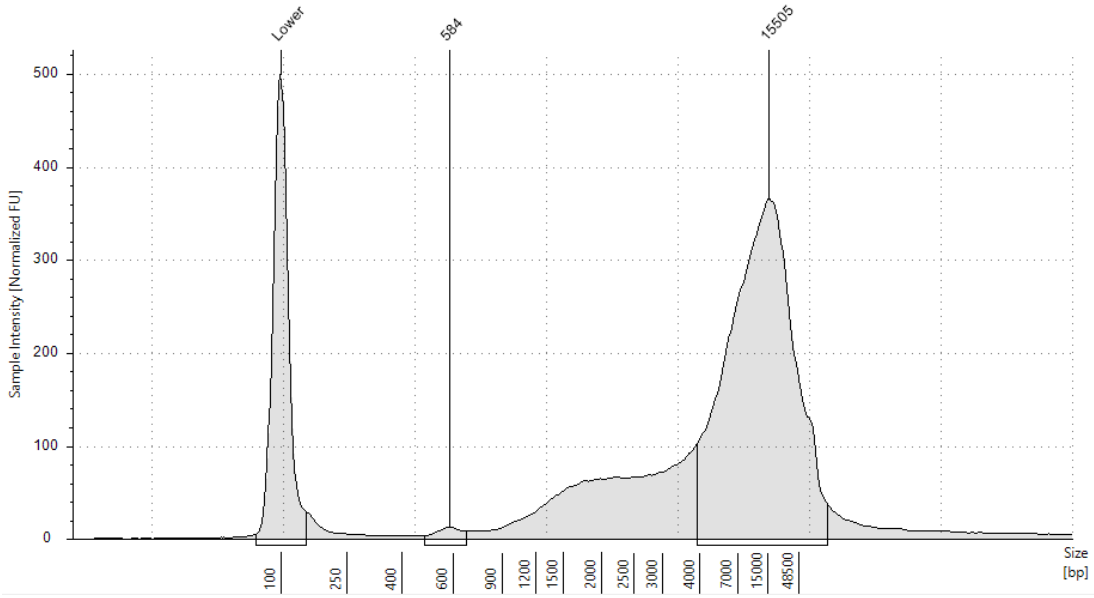

Figure 5.4: GenomeTape TapeStation trace of the Nanopore library for 165max.

The sequencing of 164fm yielded 11.6 Gb in a little over 20 hrs with a N50 on the MinION at 7.45 kb and a mean read quality of 16 (Table 5.3). This represented the most successful run. It was also the only flowcell that, after stopping, retained sufficient active pores to support subsequent isolate sequencing. 164max required the entire lifespan of the flowcell to yield 9.45 Gb. The mean read quality was comparable to the Fire Monkey prep, however the N50 was lower at 6.55 kb. There was a significant loss in DNA (ng) during the final wash step for 164max DNA. This step uses ONT's Long

Fragment buffer, which suggested that a significant amount of DNA was below 3 kb. For stool 165 neither of the DNA preps sequenced as well; both had lower N50 scores at 1.98 kb for the Fire Monkey prep and 3.15 kb for the Maxwell prep. Interestingly, while the TapeStation is not designed specifically for HMW DNA the traces above gave a pretty clear indication of how the run would turn out. 164fm showed the cleanest and largest peak and lead to the best run performance. 164fm gave a clean peak with a lower yield and lead to a good N50 but a run that required a full flowcell to yield in the 8-10 Gb range. For both 165fm and 165max a ridge in the TapeStation trace was visible before the larger peaks. This lead to a lower N50 on the MinION run and while the concentration of the 165fm and 165max were comparable to 164fm the smaller fragment size appeared to influence the speed at which the MinION could produce 8-10 Gb (Table 5.3).

Table 5.3: Basic MinION Run Information

Stool ID	DNA prep	N50 (kb)	Reads (M)	Bases (Gb)	Run time (hrs:mins)	Mean read quality
164	FM	7.45	2.51	11.6	20:10	16.0
164	Max	6.55	3.30	9.45	72:00	15.9
165	FM	1.98	5.20	7.06	67:03	19.9
165	Max	3.15	9.10	15.5	67:04	18.5

5.4.3 Long-read size and quality filtering

Prior to downstream analysis all sequencing files were filtered to remove DNA sequences below 500 bp and below a quality score of 10. A Q-score of 10 in ONT sequencing corresponds to 90% base accuracy and is commonly used as a practical minimum threshold to ensure reliable read quality while preserving enough data for meaningful analysis. Filtering at this level helped mitigate downstream errors while keeping sequencing yield reasonably high. In all samples a slight increase in N50 and read quality could be seen after filtering and as expected this came at the cost of a decrease in the number of reads (Tables 5.4-5.7).

Table 5.4: Read Statistics for Stool 164 Fire Monkey DNA Prep

	Raw	Q10 >500 bp
Mean read length	4,570.2	5,049.6
Mean read quality	16.0	16.6
Median read length	2,900.0	3,241.0
Median read quality	16.6	16.8
Number of reads	2,565,014.0	2,181,423.0
Read length N50	7,485.0	7,550.0
Total bases	11,722,734,169.0	11,015,261,568.0

Table 5.5: Read Statistics for Stool 164 Maxwell DNA Prep

	Raw	Q10-500 bp
Mean read length	2,805.9	3,729.0
Mean read quality	15.9	16.9
Median read length	1,143.0	1,985.0
Median read quality	16.5	17.1
Number of reads	3,340,127.0	2,254,060.0
Read length N50	6,562.0	6,843.0
Total bases	9,372,151,191.0	8,405,350,909.0

Table 5.6: Read Statistics for Stool 165 Fire Monkey DNA Prep

	Raw	Q10-500 bp
Mean read length	1,341.3	1,714.8
Mean read quality	19.9	21.0
Median read length	864.0	1,212.0
Median read quality	20.2	21.0
Number of reads	5,328,835.0	3,687,121.0
Read length N50	1,955.0	2,106.0
Total bases	7,147,371,778.0	6,322,727,667.0

Table 5.7: Read Statistics for Stool 165 Maxwell DNA Prep

	Raw	Q10-500 bp
Mean read length	1,616.7	2,155.4
Mean read quality	18.5	20.1
Median read length	791.0	1,202.0
Median read quality	18.7	20.3
Number of reads	9,477,082.0	6,297,924.0
Read length N50	3,058.0	3,350.0
Total bases	15,321,450,043.0	13,574,391,387.0

5.4.4 Basic fasta statistical comparison of long- and short-read *Campylobacter* MAGs

Initially, I performed binning on long-read datasets using MetaMBDG, a tool originally developed for PacBio data. However, advancements in ONT chemistry, particularly with R10 flow cells, have enabled effective support for ONT reads in MetaMBDG. Since no single tool performs optimally on both long- and short-read data, I used the pipeline MetaWrap2 for short-read assemblies. Drawing on experience from Chapter 3, I applied Kraken2 to recover reads from both long- and short-read datasets, focusing on those classified as *Campylobacter* (taxonomic ID 194), followed by targeted assembly.

For Stool 164, assemblies varied significantly depending on the DNA extraction method, sequencing technology, and assembly approach used. When using the Fire Monkey prep with long reads, MetaMBDG yielded the most cohesive assembly, producing a single contig with a total length of 1,713,954 bp. In contrast, Kraken generated nine contigs totalling 1,664,675 bp, with a N50 of 283,432 bp. Switching to short reads with the Fire Monkey prep, MetaWrap2 resulted in 86 contigs spanning 1,686,376 bp, with an N50 of 30,596 bp, indicating moderate fragmentation. Conversely, Kraken produced a higher number of contigs (241) totalling 1,614,563 bp, with an even lower N50 of 16,735 bp.

Under the Maxwell prep for long reads, MetaMBDG again excelled with a single contig assembly spanning 1,714,963 bp, similar to its performance with the Fire Monkey prep. Kraken produced five contigs totalling 1,700,530 bp, with an N50 of 479,664 bp, indicating slightly more fragmentation compared to MetaMBDG but still maintaining a high-quality assembly. In short reads under Maxwell, MetaWrap2 produced 86 contigs totalling 1,678,035 bp, with an N50 of 36,769 bp, while Kraken resulted in 377 contigs spanning 1,873,129 bp, with an N50 of 17,868 bp (Table 5.8).

Similarly, for Stool 165, the choice of prep, read type, and assembly approach significantly affected the assembly outcomes. Using the Fire Monkey prep with long reads, MetaMBDG produced 113 contigs totaling 1,371,102 bp, with an N50 of 20,583 bp, indicating moderate fragmentation. Kraken yielded 91 contigs spanning 1,201,509 bp, with a slightly higher N50 of 23,486 bp, suggesting comparable but slightly more fragmented results compared to MetaMBDG. In short reads with Fire Monkey, MetaWrap2 resulted in 224 contigs covering 1,585,898 bp, with an N50 of

9,383 bp, indicating higher fragmentation likely due to the shorter read length. Kraken, on the other hand, generated 426 contigs totalling 1,670,800 bp, with an N50 of 7,333 bp, indicating more extensive fragmentation despite a higher total length, possibly due to the inclusion of redundant or misassembled sequences.

Under the Maxwell prep for long reads, MetaMBDG produced 26 contigs spanning 1,731,358 bp, with an N50 of 280,628 bp, indicating a highly contiguous assembly with minimal fragmentation. Kraken resulted in 45 contigs totalling 1,764,205 bp, with an N50 of 76,211 bp, showing slightly more fragmentation compared to MetaMBDG but still achieving a high-quality assembly. In short reads under Maxwell, MetaWrap2 yielded 138 contigs covering 1,624,222 bp, with an N50 of 16,346 bp, while Kraken produced 274 contigs totalling 1,685,552 bp, with an N50 of 12,624 bp. GC content remained stable (~30.3–31.1%) across all assemblies (Table 5.9).

These results highlight how extraction method and sequencing strategy significantly influence assembly quality and completeness across diverse samples. They also demonstrate the clear advantage of pairing high-integrity DNA extraction with long-read sequencing and MetaMBDG, showcasing superior assembly quality and integrity.

Table 5.8: Fasta Statistics on Stool 164 MDGs Assembled Using the Different Pipelines

Prep ID	Reads	Approach	# contigs	Total length (bp)	# contigs	Largest contig (bp)	Total length (bp)	GC (%)	N50 (bp)
164fm	Long	MetaMBDG	1	1,713,954	1	1,713,954	1,713,954	30.45	1,713,954
164fm	Long	Kraken	9	1,664,675	9	357,409	1,664,675	30.49	283,432
164fm	Short	MetaWrap2	86	1,686,376	86	105,938	1,686,376	30.56	30,596
164fm	Short	Kraken	241	1,614,563	212	68,909	1,603,590	30.53	16,735
164max	Long	MetaMBDG	1	1,714,963	1	1,714,963	1,714,963	30.45	1,714,963
164max	Long	Kraken	5	1,700,530	5	653,010	1,700,530	30.44	479,664
164max	Short	MetaWrap2	86	1,678,035	86	138,721	1,678,035	30.57	36,769
164max	Short	Kraken	377	1,873,129	295	105,288	1,841,813	30.24	17,868

Table 5.9: Fasta statistics on Stool 165 MDGs assembled using the different pipelines

Stool ID	Reads	Approach	# contigs	Total length (bp)	# contigs	Largest contig (bp)	Total length (bp)	GC (%)	N50 (bp)
165fm	Long	MetaMBDG	113	1,371,102	113	76,249	1,371,102	30.36	20,583
165fm	Long	Kraken	91	1,201,509	91	50,746	1,201,509	30.49	23,486
165fm	Short	MetaWrap2	224	1,585,898	224	36,336	1,585,898	30.45	9,383
165fm	Short	Kraken	426	1,670,800	389	32,157	1,656,380	30.42	7,333
165max	Long	MetaMBDG	26	1,731,358	26	511,772	1,731,358	31.09	280,628
165max	Long	Kraken	45	1,764,205	45	173,811	1,764,205	30.33	76,211
165max	Short	MetaWrap2	138	1,624,222	138	68,279	1,624,222	30.39	16,346
165max	Short	Kraken	274	1,685,552	252	41,463	1,676,847	30.34	12,624

5.4.5 MAG completeness

The binning results using long reads showed clear differences in genome quality across stool samples and DNA extraction methods. For Stool 164, both the Fire Monkey and Maxwell preps produced high-quality bins using SemiBin, with 99.85% completeness, 0.23% contamination, and no strain heterogeneity, indicating near complete and clean genome reconstructions. In contrast, Stool 165 yielded more variable and lower-quality bins. The Fire Monkey prep resulted in a bin with 74.37% completeness, 0.50% contamination, and 20.00% strain heterogeneity, while the Maxwell prep performed better with 90.85% completeness and lower contamination (0.32%), though it still exhibited 20.00% strain heterogeneity. These results reinforce earlier findings that DNA quality and sequencing depth significantly affect bin quality, with Maxwell extractions producing more complete and reliable bins, particularly under more challenging conditions.

The binning results using short reads indicate that all assemblies achieved high completeness (≥96.00%) and low contamination (<2.00%), suggesting overall strong recovery of target genomes across all preps and samples. However, strain heterogeneity was substantial in all cases, ranging from 66.67% to 83.33%, indicating the presence of multiple closely related strains within each bin. Specifically, for 164fm and 164max produced bins with >98.50% completeness, though strain heterogeneity was high (66.67% and 75.00%, respectively). Similarly, for Stool 165, both preps produced slightly lower completeness (96.00–97.60%), with strain heterogeneity also exceeding 66%. These findings suggest that while genome reconstruction was effective in terms of completeness and contamination, strain-level diversity remains a major challenge in these metagenomic bins.

The results from the Kraken classification and assembly of recovered reads approach showed strong performance for stool 164 across both preps and sequencing types. Completeness was ≥97.60% in all cases, with zero contamination and strain heterogeneity in long-read assemblies, and only minimal issues in short reads. Notably, the Maxwell short-read assembly achieved 99.96% completeness, though with elevated contamination (5.10%) and high strain heterogeneity (92.00%), suggesting possible over-assembly or misclassification. For stool 165, results were

more mixed. The Fire Monkey long-read assembly had low completeness (62.32%) and high strain heterogeneity (66.67%), indicating poor genome recovery and strain-level complexity. In contrast, the Maxwell long-read assembly was nearly complete (98.51%) but had moderate contamination (2.89%) and very high strain heterogeneity (93.33%). Short-read assemblies for stool 165 performed well overall, with >98.8% completeness and low contamination**,** particularly in165max, which also showed 0% strain heterogeneity, the best overall result for this stool. These findings highlight that Kraken-based read recovery can yield high-quality bins, especially when paired with Maxwell extraction and short-read data, although strain heterogeneity remains a challenge, particularly in complex or low-quality samples (Tables 5.10-5.12).

Table 5.10: CheckM Results on MetaMBDG *Campylobacter* Bins Using Marker Lineage *Campylobacter* (UID3076)

Prep ID	Bin ID	Completeness (%)	Contamination (%)	Strain heterogeneity
164fm	SemiBin_1009	99.85	0.23	0.00
164max	SemiBin_490	99.85	0.23	0.00
165fm	SemiBin_1038	74.37	0.50	20.00
165max	SemiBin_717	90.85	0.32	20.00

Table 5.11: CheckM Results on MetaWrap2 *Campylobacter* Bins Using Marker Lineage *Campylobacter* (UID3076)

Prep ID	Bin ID	Completeness (%)	Contamination (%)	Strain heterogeneity
164fm	bin20_270	99.21	1.71	66.67
164max	bin22_409	98.58	0.8	75.00
165fm	bin2_110	96.02	0.74	83.33
165max	bin4_139	97.64	0.76	66.67

Table 5.12: CheckM Results on *Campylobacter* Reads Recovered Using Kraken2

Prep ID	Reads	Completeness	Contamination	Strain heterogeneity
164fm	Long	97.68	0.00	0.00
164max	Long	99.58	0.00	0.00
165fm	Long	62.32	0.44	66.67
165max	Long	98.51	2.89	93.33
164fm	Short	99.03	0.98	16.67
164max	Short	99.96	5.10	92.00
165fm	Short	98.88	1.69	11.11
165max	Short	99.64	1.33	0.00

5.4.6 *Campylobacter* isolates

To establish reference standards for comparison with the MAG-derived typing results, *Campylobacter* was isolated from the two stool samples, and hybrid genome assemblies were generated to determine isolate-level typing information.

5.4.6.1 164

For Stool 164 all six *Campylobacter* isolates were identified as *C. jejuni* ST22 and carried a single antimicrobial resistance determinant, $bla_{\text{OXA-592}}$.

5.4.6.2 165

For stool 165 all six *Campylobacter* isolates were all confirmed to be *C. jejuni* ST5136, with an AMR profile consisting of $bla_{\text{OXA-592}}$, *50S_L22_A103V*, *tet*(O), *gyrA*_T86I.

5.4.7 Mapping metagenomic reads to isolates

To determine sequencing coverage scores for *Campylobacter* in the stool samples, metagenomic sequencing reads were aligned to a hybrid reference genome assembly. Specifically, isolate 164-5 was used for stool 164, while stool 165 was mapped against isolate 165-3. For stool sample 164, both long and short read approaches resulted in high coverage of the target *Campylobacter* genome (38.70-92.14x) (Tables 5.13 & 5.14). For the long read sequencing of stool sample 164 100% breadth of coverage was achieved with both DNA prep methods.

Table 5.13: Mapping of Metagenome Sequencing Long Reads to *Campylobacter* Isolate

Prep ID	Sequencing (Gb)	Coverage	Coverage per 10 Gb	Breadth
164fm	11.90	57.11	48.00	100.00
164max	10.00	92.14	92.14	100.00
165fm	6.70	4.71	7.03	98.67
165max	14.00	14.12	11.86	100.00

Table 5.14: Mapping of Metagenome Sequencing Short Reads to *Campylobacter* Isolate

Prep ID	Sequencing (Gb)	Coverage	Coverage per 10 Gb	Breadth
164fm	8.98	40.47	45.06	99.95
164max	11.57	44.78	38.70	99.96
165fm	13.53	13.44	9.93	99.80
165max	12.03	16.03	13.33	99.91

5.4.8 MAGs versus isolate typing

For the stool 164 MAGs, sequence typing results showed a clear match to the reference isolates (Table 5.15). Only one allele was missed in the Fire Monkey DNA extraction when using the short-read binning approach. In contrast, the results for Stool 165 were more striking: while all short-read approaches yielded complete STs, none of the long-read approaches recovered a full ST profile. Notably, 165max outperformed 165fm, likely due to higher sequencing yield and larger fragment sizes, which improved genome recovery and typing accuracy (Table 5.16).

Table 5.15: MLST Results for Stool 164 for Fire Monkey and Maxwell Preps Sequenced by MinION (Long) and Illumina Pair-end 150 bp (Short) with Assemblies Created by Binning and Read Classification Approaches

Stool ID	Reads	Approach	ST	*aspA*	*glnA*	*gltA*	*glyA*	*pgm*	*tkt*	*uncA*
164fm	Short	Binning	-	1	3	6	4	-	3	3
164max	Short	Binning	22	1	3	6	4	3	3	3
164fm	Short	Read classification	22	1	3	6	4	3	3	3
164max	Short	Read classification	22	1	3	6	4	3	3	3
164fm	Long	Binning	22	1	3	6	4	3	3	3
164max	Long	Binning	22	1	3	6	4	3	3	3
164fm	Long	Read classification	22	1	3	6	4	3	3	3
164max	Long	Read classification	22	1	3	6	4	3	3	3

Table 5.16: MLST Results for Stool 165 for Fire Monkey and Maxwell Preps Sequenced by MinION (Long) and Illumina Pair-end 150 bp (Short) with Assemblies Created by Binning and Read Classification Approaches

Stool ID	Reads	Approach	ST	*aspA*	*glnA*	*gltA*	*glyA*	*pgm*	*tkt*	*uncA*
165fm	Short	Binning	5136	24	2	2	2	10	3	3
165max	Short	Binning	5136	24	2	2	2	10	3	3
165fm	Short	Read classification	5136	24	2	2	2	10	3	3
165max	Short	Read classification	5136	24	2	2	2	10	3	3
165fm	Long	Binning	-	24	-	2	2	-	738?	-
165max	Long	Binning	-	24	2	2	747?	10	3	3
165fm	Long	Read classification	-	24	-	2	2	-	738?	-
165max	Long	Read classification	-	24	2	2	719?	10	3	3

5.4.9 MAGs versus isolate antimicrobial determinants

The detection of AMR in the MAGs following a similar pattern to the ST results. For stool 164 all approaches yielded a complete AMR profiles in line with the 164 isolates (Table 5.17). For the 165 MAGs the short read approaches identified 3 out of the 4 AMR determinants. Interestingly, *tet*(O) was not identified in short read-derived MAGs. This was also the case in chapter 3 with *tet* gene variants identified in isolate sequencing but not in MAGs. Mixed results were present for the long-read approaches for stool 165. For the 165max the binning approach did identify *tet*(O), however it failed to identify the gyrase mutation. Read classification from long reads from 165fm was the poorest performing, only identifying the 50S mutation L22_A103V (Table 5.18).

Table 5.17: AMR Detection Results for Stool 164 for Fire Monkey and Maxwell Preps Sequenced by MinION (Long) and Illumina Pair-end 150 bp (Short) with Assemblies Created by Binning and Read Classification Approaches

Stool ID	Reads	Approach	Beta-lactamase
164fm	Short	Binning	$bla_{OXA-592}$
164max	Short	Binning	$bla_{OXA-592}$
164fm	Short	Read classification	$bla_{OXA-592}$
164max	Short	Read classification	$bla_{OXA-592}$
164fm	Long	Binning	$bla_{OXA-592}$
164max	Long	Binning	$bla_{OXA-592}$
164fm	Long	Read classification	$bla_{OXA-592}$
164max	Long	Read classification	$bla_{OXA-592}$

The expected AMR profile based off isolate sequencing was $bla_{OXA-592}$.

Table 5.18: AMR Detection Results for Stool 165 for Fire Monkey and Maxwell Preps Sequenced by MinION (Long) and Illumina Pair-end 150 bp (Short) with Assemblies Created by Binning and Read Classification Approaches

Stool ID	Reads	Approach	Beta-lactamase	Quinolone	Macrolide	Tetracycline
165fm	Short	Binning	*bla*$_{\text{OXA-592}}$	*gyrA*_T86I	*50S_L22_A103V*	-
165max	Short	Binning	*bla*$_{\text{OXA-592}}$	*gyrA*_T86I	*50S_L22_A103V*	-
165fm	Short	Read classification	*bla*$_{\text{OXA-592}}$	*gyrA*_T86I	*50S_L22_A103V*	-
165max	Short	Read classification	*bla*$_{\text{OXA-592}}$	*gyrA*_T86I	*50S_L22_A103V*	-
165fm	Long	Binning	*bla*$_{\text{OXA-592}}$	*gyrA*_T86I	*50S_L22_A103V*	-
165max	Long	Binning	*bla*$_{\text{OXA-592}}$	-	*50S_L22_A103V*	*tet*(*O*)
165fm	Long	Read classification	-	-	*50S_L22_A103V*	-
165max	Long	Read classification	*bla*$_{\text{OXA-592}}$	*gyrA*_T86I	*50S_L22_A103V*	-

The expected AMR profile based off isolate sequencing was *bla*$_{\text{OXA-592}}$, *gyrA*_T86I, *50S_L22_A103V*, and *tet*(O). When "–" is present in Table, expected AMR was missing.

5.4.10 Tetracycline resistance determinants

AMR determinates were identified using AbritAMR. The *tet*(O) gene was identified in the hybrid isolate assemblies recovered from stool 165 when creating a profile to screen for in the MAGs. This gene was only identified in a MAG by AbritAMR when using the long-read binning approach from 165max. This was a significant finding and potentially a strength of long-read sequencing and the Maxwell DNA prep. To explore this, I manually screened genome annotations of isolate 165-3 and all MAGs created from the stool 165. Using the sequence data for isolate 165-3, three genome assemblies were screened. A hybrid, a long-read only, and a short read only assembly. A *tet*(O) gene was manually identified at the expected genome location in all three 165-3 isolate assemblies (Figure 5.5). In the MAGs a *tet*(O) gene was manually identified when using long-read binning approaches for 165fm and 165max. For 165max both binning and Kraken based read recovery approaches resulted in the identification of *tet*(O). I was unable to locate the *tet*(O) gene in any of the short-read MDGs (Table 5.19).

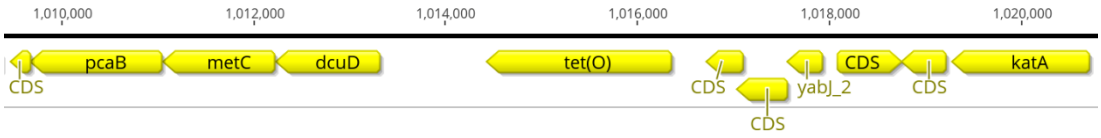

Figure 5.5: Location of the *tet(O)* gene in the genome of isolate 165-3.

Table 5.19: Manual and AbritAMR Identification of tet(O) in MDGs and Isolate Sequencing

Stool ID	Reads	Approach	*tet*(O) manual identification	*tet*(O) AbritAMR
165fm	Short	MDG-Binning	No	No
165max	Short	MDG-Binning	No	No
165fm	Short	MDG-Read classification	No	No
165max	Short	MDG-Read classification	No	No
165fm	Long	MDG-Binning	Yes	No
165max	Long	MDG-Binning	Yes	Yes
165fm	Long	MDG-Read classification	No	No
165max	Long	MDG-Read classification	Yes	No
165	Long	Isolate sequencing	Yes	Yes
165	Short	Isolate sequencing	Yes	Yes
165	Hybrid	Isolate sequencing	Yes	Yes

5.4.11 MAGs versus isolate single nucleotide polymorphisms

For stool 164 long-read binning delivered the highest alignment accuracy and coverage, with minimal unaligned bases, zero variant calls, and near-complete agreement with the hybrid reference. In contrast, short-read classification approaches were the most error-prone, showing elevated unaligned regions, variant calls, and low coverage zones particularly for the Fire Monkey prep. These results highlight the superiority of long-read binning for generating MAGs closely matching isolate genomes, and show that read classification, especially with short reads, may introduce substantial noise in variant analyses (Table 5.20).

For stool 165, Maxwell DNA extractions outperformed Fire Monkey extractions, especially in long-read assemblies. Long-read binning with Maxwell delivered the closest match to the reference, with near-complete alignment, very few variants, and minimal low coverage. In contrast, Fire Monkey long-read data was highly fragmented, with extensive unaligned regions and elevated variant and heterozygosity rates, suggesting lower sequencing quality or strain complexity. Among short-read approaches, binning was generally more reliable than read classification, though both showed elevated low coverage and modest levels of unaligned bases compared to long-read strategies (Table 5.21).

Table 5.20: Single Nucleotide Polymorphism Analysis Results for Stool 164 for Fire Monkey and Maxwell Preps Sequenced by MinION (Long) and Illumina Pair-end 150 bp (Short) with Assemblies Created by Binning and Read Classification Approaches

Stool ID	Reads	Approach	LENGTH	ALIGNED	UNALIGNED	VARIANT	HET	LOWCOV
164fm	Short	Binning	1,713,982	1,650,101	49,825	8	1	14,055
164max	Short	Binning	1,713,982	1,622,258	77,922	15	0	13,802
164fm	Short	Read classification	1,713,982	1,552,488	113,360	41	0	48,134
164max	Short	Read classification	1,713,982	1,579,550	103,025	1,785	1,641	29,766
164fm	Long	Binning	1,713,982	1,685,794	22,710	0	0	5,478
164max	Long	Binning	1,713,982	1,684,959	22,616	0	0	6,407
164fm	Long	Read classification	1,713,982	1,635,916	71,875	379	787	5,404
164max	Long	Read classification	1,713,982	1,662,730	46,103	45	34	5,115
Reference	Hybrid	Isolate assembly	1,713,982	1,713,982	0	0	0	0

LENGTH: Length of the sequence alignment in base pairs, indicating the span of the genomic region analysed. ALIGNED: Number of sequences that align perfectly with the reference sequence at the specified position. UNALIGNED: Number of sequences that do not align with the reference sequence at the given genomic position. VARIANT: Number of sequences that exhibit variants compared to the reference sequence at the analysed position. HET: Number of heterozygous variants detected, indicating the presence of two different alleles at the genomic position. LOWCOV: Number of positions with low sequencing coverage, impacting the reliability of variant calls due to insufficient read depth.

Table 5.21: Single Nucleotide Polymorphism Analysis Results for Stool 165 for Fire Monkey and Maxwell Preps Sequenced by MinION (Long) and Illumina Pair-end 150 bp (Short) with Assemblies Created by Binning and Read Classification Approaches

Stool ID	Reads	Approach	LENGTH	ALIGNED	UNALIGNED	VARIANT	HET	LOWCOV
165fm	Short	Binning	1,743,227	1,521,734	175,661	64	0	45,832
165max	Short	Binning	1,743,227	1,584,184	130,561	29	1	28,481
165fm	Short	Read classification	1,743,227	1,561,535	94,992	49	0	86,700
165max	Short	Read classification	1,743,227	1,607,674	79,772	23	0	55,781
165fm	Long	Binning	1,743,227	1,206,148	508,114	482	435	28,530
165max	Long	Binning	1,743,227	1,513,078	216,196	4	121	13,832
165fm	Long	Read classification	1,743,227	1,105,316	612,684	677	1,333	23,894
165max	Long	Read classification	1,743,227	1,643,995	84,339	7	1,953	12,940
Reference	Hybrid	Isolate assembly	1,743,227	1,743,227	0	0	0	0

LENGTH: Length of the sequence alignment in base pairs, indicating the span of the genomic region analysed. ALIGNED: Number of sequences that align perfectly with the reference sequence at the specified position. UNALIGNED: Number of sequences that do not align with the reference sequence at the given genomic position. VARIANT: Number of sequences that exhibit variants compared to the reference sequence at the analysed position. HET: Number of heterozygous variants detected, indicating the presence of two different alleles at the genomic position. LOWCOV: Number of positions with low sequencing coverage, impacting the reliability of variant calls due to insufficient read depth.

5.4.12 Community composition

In a final step I looked at community composition between the two DNA preps and for the long- and short-read datasets. For stool 164 the most abundant genus identified with 164fm was *Parabacteroides* at 28.22% for long reads and 29.86% for short reads. For 164max the most abundant genus was *Bacteroides* at 42.7% for long reads and 41.67% for short reads. The relative abundance of *Campylobacter* was markedly higher in long-read sequencing datasets, with the 164max long-reads showing the strongest recovery (12.58%). Short-read approaches consistently underestimated *Campylobacter* abundance, compared to long-read approaches. *Phocaeicola* showed notably higher relative abundance in short-read datasets, particularly in the 164max short-read prep, where it reached 16.41%, the highest across all conditions. In contrast, *Phocaeicola* abundance was lower in long-read datasets, with 8.99% in 164max long reads and 7.21% in 164fm long reads (Fig.5.6 and Table 5.22).

Table 5.22: Relative Abundance of the Top 15 Genera for Stool 164 for Fire Monkey and Maxwell Preps Sequenced by MinION (Long) and Illumina Pair-end 150 bp (Short)

Taxon	164fm Long	164fm Short	164max Long	164max Short
Parabacteroides	28.22	29.86	19.25	20.04
Bacteroides	25.77	19.37	42.70	41.67
Faecalibacterium	10.91	15.10	4.61	4.42
Klebsiella	9.13	6.47	2.28	5.16
Campylobacter	8.30	3.90	12.58	2.98
Phocaeicola	7.21	11.83	8.99	16.41
Veillonella	2.37	3.55	1.21	1.73
Escherichia	2.29	3.32	3.62	3.49
Alistipes	1.59	1.89	1.72	2.26
Streptococcus	1.42	1.59	0.42	0.32
Blautia	0.78	0.83	0.29	0.13
Roseburia	0.55	0.39	0.22	0.10
Haemophilus	0.42	0.45	0.26	0.27
Fusobacterium	0.33	0.62	0.63	0.20
Clostridium	0.13	0.17	0.34	0.07

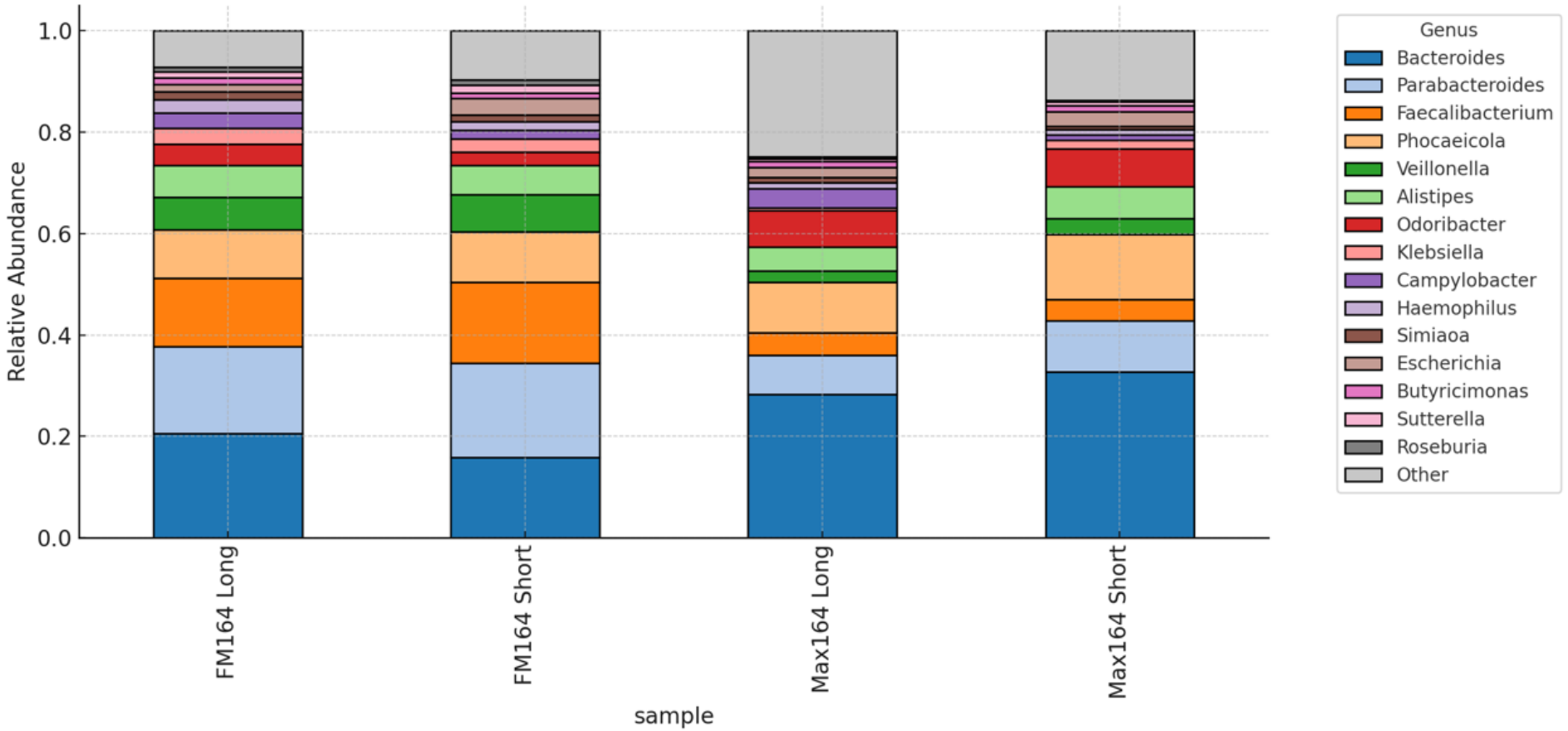

Figure 5.6: Relative abundance of the Top 15 Genera for stool 164 for Fire Monkey and Maxwell preps sequenced by MinION (Long) and Illumina Pair-end 150bp (Short).

For stool 165 a high proportion of the reads were human even after the host depletion was carried during the DNA extraction process (Fig.5. 7). I have seen throughout this PhD project that the host depletion can be temperamental when there is a high load of human DNA in the stool sample. The high human DNA load also remained in the short read data sets even after *in silico* human read removal. To counter this the human reads were removed from the Kraken reports and the relative abundances were recalculated. *Campylobacter* dominated the microbial profile across all extraction and sequencing approaches, with long-read data particularly from the Maxwell prep yielding the highest abundance. The relative stability of other abundant taxa (*Phocaeicola, Bacteroides, and Streptococcus*) across conditions suggests consistent detection (Table 5.23 & Figure 5.8).

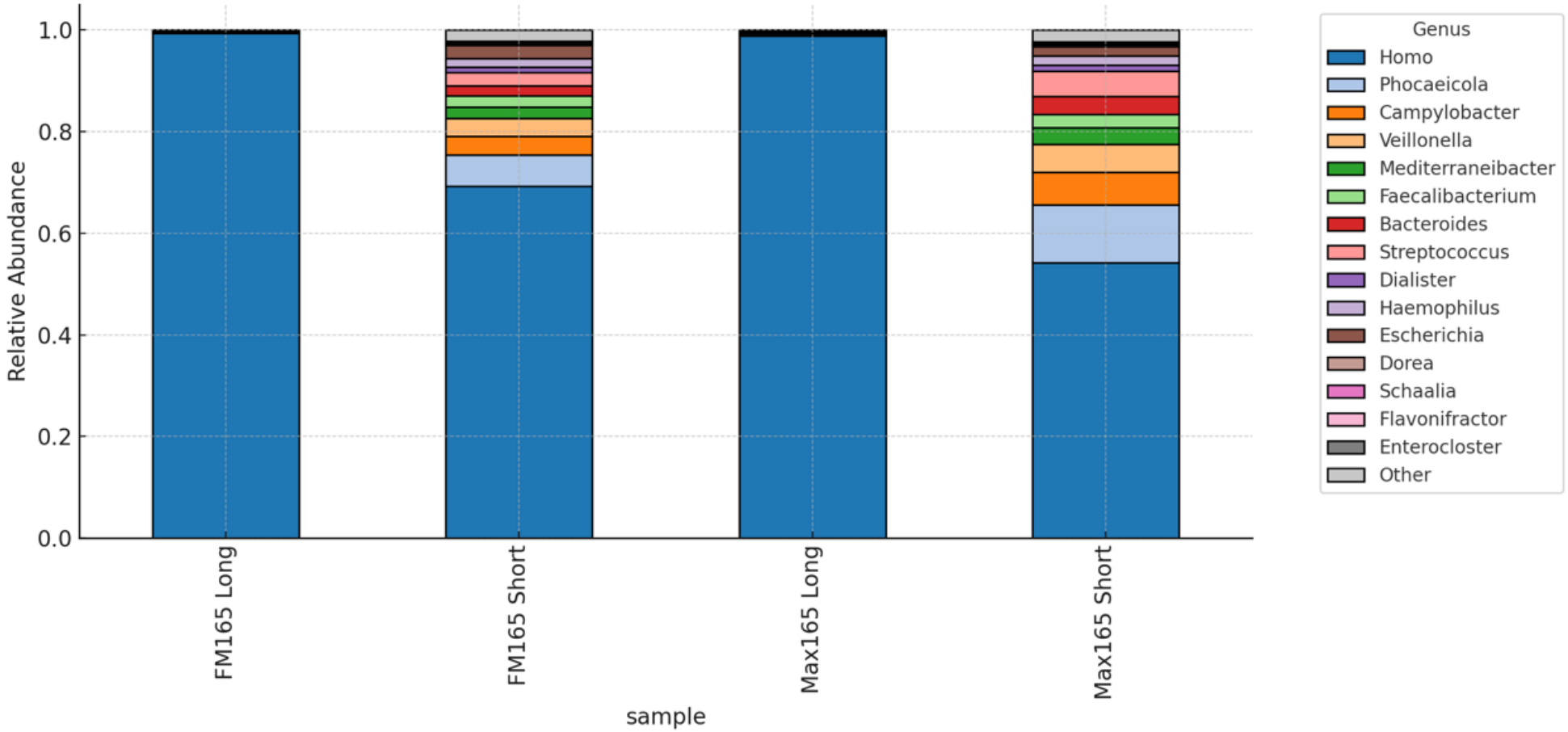

Figure 5.7: Relative abundance of the Top 15 Genera for stool 165 for Fire Monkey and Maxwell preps sequenced by Oxford Nanopore MinION (Long) and Illumina Pair-end 150bp (Short) before removal of *Homo* genus.

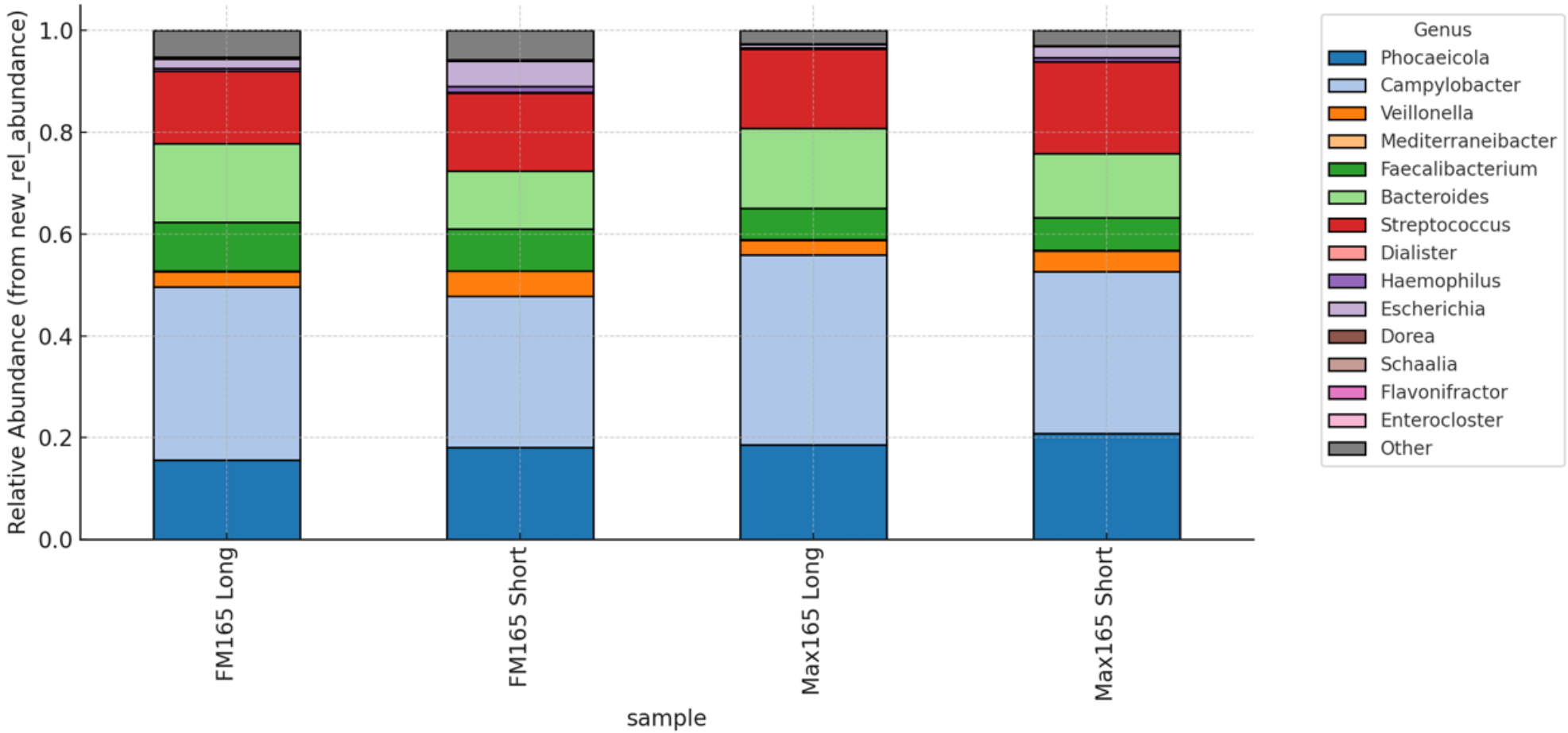

Figure 5.8: Relative abundance of the Top 15 Genera for stool 165 for Fire Monkey and Maxwell preps sequenced by Oxford Nanopore MinION (Long) and Illumina Pair-end 150bp (Short) after removal of *Homo* genus

Table 5.23: Relative Abundance of the Top 15 Genera for Stool 165 for Fire Monkey and Maxwell Preps Sequenced by MinION (Long) and Illumina Pair-end 150 bp (Short) After Removal of *Homo* genus

Taxon	165fm Long	165fm Short	165max Long	165max Short
Campylobacter	33.99	29.75	37.22	33.99
Phocaeicola	15.57	18.09	18.67	15.57
Bacteroides	15.53	11.36	15.65	15.53
Streptococcus	14.18	15.36	15.54	14.18
Faecalibacterium	9.50	8.26	6.18	9.50
Veillonella	3.07	4.87	2.91	3.07
Escherichia	1.86	4.91	0.67	1.86
Klebsiella	1.04	2.01	0.26	1.04
Enterobacter	1.04	1.33	0.21	1.04
Blautia	0.70	0.36	0.62	0.70
Haemophilus	0.55	1.29	0.32	0.55
Fusobacterium	0.52	0.47	0.68	0.52
Neisseria	0.30	0.10	0.07	0.30
Actinomyces	0.28	0.41	0.12	0.28
Schaalia	0.28	0.22	0.10	0.28

5.4.13 qPCR

The four samples were assessed by qPCR to detect the presence of *Campylobacter* DNA using the *cadF* gene, alongside human DNA detection using the RNA Polymerase II (POLR2A) gene. The results indicated elevated levels of human DNA (Cp ~25) in the 165 stool sample (Table 5.24), as corroborated by the community composition analysis (Figure 5.7). A high Cp value (~31-33) for POLR2A in stool sample 164 indicated minimal presence of human DNA (Table 5.24), consistent with the community composition analysis, where human DNA did not interfere with the results.

Table 5.24: *POLR2A* qPCR Values for Stool DNA Preps and Human Control DNA

Sample	DNA input (ng)	Human Cp
164fm stool DNA	23.00	32.73
164max stool DNA	22.00	31.12
165fm stool DNA	20.00	25.13
165max stool DNA	23.00	25.37
Human control DNA	20.00	25.77
Human control DNA	2.00	29.33
Human control DNA	0.02	32.93

For *Campylobacter* detection the level between the preps and stool samples were similar apart from the 164 stool DNA from the maxwell prep which by qPCR quantification was higher with a Cp at ~24 versus ~27 (Table 5.25).

Table 5.25: *cadF* qPCR Values for Stool DNA Preps and *Campylobacter* Control DNA

Sample	DNA input (ng)	cadF Cp
164fm stool DNA	23.00	26.82
164max stool DNA	22.00	23.96
165fm stool DNA	20.00	27.14
165max stool DNA	23.00	27.53
Campylobacter control DNA	2.00	21.80
Campylobacter control DNA	0.20	25.77
Campylobacter control DNA	0.02	29.91

5.5 Discussion

In this study, I evaluated two DNA extraction protocols for *Campylobacter* detection in human stool: the Fire Monkey HMW stool protocol (fm; Chapter 2) and the Promega Maxwell RSC Fecal Microbiome kit (max). One of the most striking findings from this study is that when optimal long-read sequencing is achieved from stool-derived DNA, it is possible to recover complete *Campylobacter* genomes and perform high-resolution typing, including single nucleotide polymorphism analysis. However, this capability declines sharply as sequencing quality diminishes, most notably reflected in reduced sequencing read N50 values. While an exact threshold for successful genome reconstruction and SNP-level resolution remains unclear, assemblies generated from datasets with read N50 values in the range of 6.8–7.5 kb yielded one single contig MAGs suitable for detailed typing. In contrast, assemblies from reads with an N50 of ~3.5 kb showed reduced completeness, and this decline was even more pronounced at N50 values near 2 kb, which manifest as assemblies in many contigs (91-113 contigs).

This progressive decline in assembly quality was observed in all downstream analyses, including species classification, ST, AMR profiling, and SNP detection. In single-contig assemblies, genes remain intact rather than fragmented across contigs, which improves the performance of bioinformatic tools that depend on DNA sequence

databases. Correctly assembled genes are more likely to produce confident matches in these databases. Achieving a single-contig genome is often hindered by ribosomal RNA operons, repetitive 5–6 kb sequences that can appear in multiple copies throughout bacterial genomes (Koren & Phillippy, 2015; Martins et al., 2020). A more conservative target would be to aim for a read N50 ~20 kb, or if knowledge of the target genomes is available as large as the largest repeat (Schmid et al., 2018; Wick et al., 2023).

Assembly becomes even more challenging in metagenomic datasets, where strain-level diversity compounds the difficulty of resolving repetitive DNA. Increasing the read N50 offers one solution, as longer reads are more likely to span repeats or conserved regions and anchor them within unique, strain-specific flanking sequences. Oxford Nanopore long-read metagenomics has demonstrated this in practice, yielding complete, circularised genomes of repeat-rich organisms such as *Prevotella copri,* particularly when the species was abundant in the sample (Moss et al., 2020). This was achieved with a read N50 of 3,030 bp, a longest read of 115,448 bp, and 765x coverage of the organism. In the same study, the lowest coverage at which a complete circular genome was recovered was 75x for an *Oscillibacter* species, using a DNA from a stool sample with a read N50 of 4,654 bp and a longest read of 133,658 bp (Moss et al., 2020). By contrast, short-read assemblies of the same samples were highly fragmented (contig N50 on the order of tens of kb). Notably, the *P. copri* genome was finished in one contig with ONT reads, whereas even >4,800x coverage of short reads never exceeded a 130 kb N50 for this genome (Moss et al., 2020). This is consistent with my findings: achieving coverage exceeding 57x resulted in a single contig with 100% coverage across the entire *Campylobacter* genome using long reads, whereas genomes assembled from short reads appeared fragmented with over 80 contigs.

A study by Djeghout *et al.,* reported when MAGs created from direct short read sequencing of stool reach a completeness over 60%, 72% of the MAGs yield complete ST information, sequencing typing failed below a genome complete of 60%. The study also reported that >5% coverage of the genome led to a 74% success rate in obtaining a complete ST (Bilal Djeghout et al., 2024). The short-read sequencing of the two samples in this chapter met these criteria a full ST was identified. The short-read binning approach for 164fm failed to yield a complete ST, however the read recovery

approach did obtain the complete ST. These metrics do not appear to apply to long-read sequencing as for stool 165 no complete ST was obtained with genome completeness ranging from 74.37-97.64% and genome coverage range from 4.71-16.03%. For these samples I estimated breadth of coverage to be between 98.67-99.99%. To me this suggests that quality of ONT long-read sequencing with R10 is still an issue and higher coverages are required to obtain a complete ST. Using stool 164 as a guide a coverage of 40x yield a complete ST, however I suspect coverage values in the 20x-25x would suffice as the software was one allele out of seven short of a full sequence typing when the coverage was 14-16%. In the study (Bilal Djeghout et al., 2024) a single metagenomic samples was screening for *Campylobacter* using direct ONT long-read sequencing of stool testing adaptative sampling version standard. For adaptive sequencing a full ST was obtained at 7x coverage, and an incomplete ST was obtained using the standard approach at 5x. The studied used a different approach to recover *Campylobacter* reads where the metagenomic read were mapped to a database of 602 *Campylobacter* genus sequences including plasmids. This approach may offer enhanced read recovery but is very much limited by the quality of the database and may miss novel *Campylobacter* sequences or recently acquired mobile element, for which *do novo* assembly would be better. A key issue with *do novo* assembly from metagenomic sequencing is plasmids are not associated with any genome, the origin genome is difficult to trace (Antipov et al., 2019; Krawczyk et al., 2018). Therefore a combination of *de novo* and bespoke database driven approaches my result in the most comprehensive characterisation of *Campylobacter* when directly sequencing from complex samples such as stool.

An intriguing result emerged from sequencing stool 165: short-read approaches surpassed long-read methods in MLST typing and AMR determinant detection. This finding is particularly fascinating for two reasons: firstly, despite producing longer contigs, long-read sequencing underperformed compared to short reads. Secondly, the quality of long-read data from stool 165 exceeded that of the long-read data for stool 164. A key variable to consider here is genome coverage. For stool 164 both long- and short-read approaches exceed 40x, a value that is cited as the lower end of the acceptable range for ISO-certified isolate genomics for surveillance of antimicrobial resistance (Sherry et al., 2023). In this scenario short reads provided accurate AMR information, the sequencing of the Maxwell prep gave accurate MLST information,

however the sequencing of the Fire Monkey prep missed one of seven MLST locus resulting in no full ST. Long-read sequencing resulted in full AMR and MLST information, and impressively core SNP analysis with 0 SNPs compared to the hybrid isolate reference. For stool 165 the coverage metrics should reduced coverage failing below 40x, with values between 4.7x-16x. Under these conditions the short reads perform favourably achieved complete ST information and almost complete AMR profiles. The short reads missed *tet*(O) but manged to capture *blaOXA-193*, *gyrA*_T86I, and *50S_L22_A103V*. In contrast the long-read sequencing approaches struggled at low coverage failed to yield complete ST or AMR profiles. The stool 165 sequencing of the Maxwell prep was however close to complete information, one of seven alleles was missed in MLST typing and the full AMR profile could be observed if the information from the binning and read recovery approach were combined. Taken together these results suggest 14x coverage of the target pathogen using long read or short read sequencing is right on the limit of complete MLST and AMR detection.

Although long-read MAGs generally had fewer contigs, they were more prone to sequence inaccuracies, particularly in conserved MLST loci. Tools like Galaxy's MLST require near-exact allele matches; even minor base errors, frameshifts, or assembly artefacts within target loci can prevent allele recognition and lead to incomplete STs. In contrast, despite being more fragmented, the short-read assemblies had much higher per-base accuracy, which likely preserved the integrity of MLST genes and enabled complete allele calls. High strain heterogeneity, particularly in stool 165, may have further complicated long-read assemblies, leading to chimeric or collapsed loci that obscured correct allele reconstruction. These findings underscore that, in the context of MLST from MAGs, sequence accuracy is more critical than contig number or assembly continuity, and that short-read data can outperform long-read assemblies for strain typing when long-read polishing or depth is suboptimal.

The tetracycline resistance gene *tet*(O) was uniquely detected using AbriTAMR in the Maxwell prep, leveraging long-read sequencing and a binning strategy for assembly. This gene is notable for its ability, alongside similar tetracycline resistance genes responsible for ribosomal protection proteins, to undergo recombination, forming mosaic genes (Hormeño et al., 2020b). Throughout the project, AbriTAMR encountered challenges in identifying *tet*(O) within MAGs where its presence was expected based

on isolate sequencing. Manual inspection revealed database issues in the long-read MAGs and structural issues (region split across contigs or missing) in the short-read MAGs affecting gene identification. The binning approach proved most effective in achieving a complete *tet*(O) gene sequence using long reads.

A critical step in long-read sequencing of DNA from human stool is ensuring the that the DNA is pure to the textbook standards of an A260/A280 ratio of ~1.8 and an A260/A230 ratio between 2.0 and 2.2. In the development of the Fire Monkey stool protocol some early sequencing attempts were made with heartbreakingly poor results. This I now account to DNA being outside of the above-mentioned ratios. Micrograms of DNA goes into a Nanopore library prep and this increases the chances of contaminants getting into the flowcell. Care needs to be taken during all the bead washing steps and I recommend resuspending SPRI beads in ethanol once the DNA is attached rather than pipetting ethanol over the beads on the magnet. ONT sequencing Q Scores are still low even though they are improving so every precaution needs to be made when sequencing DNA from challenging inhibitory laded sample types.

Tools such as qPCR and TapeStation play crucial roles in identifying samples likely to produce high-quality MAGs with comprehensive typing information. Elevated levels of host DNA in a sample can diminish bacterial content coverage, reducing typing success rates. Using qPCR for a human gene is effective at identifying if your host depletion protocol has been successful. For instance, stool 165 exhibited higher human DNA levels based on qPCR, resulting in reduced *Campylobacter* coverage and complications in community composition analysis; excess human reads necessitated removal, and composition values required recalibration. In retrospect, qPCR results suggested that re-extracting DNA from the stool sample with a higher saponin percentage would have been beneficial. While TapeStation tends to overestimate DNA size, it remains valuable for detecting low molecular weight DNA in extractions. The Read N50 metric is critical for achieving single-contig assemblies; detection of low molecular weight DNA prompts decisions on re-extraction or size selection prior to sequencing. While these practices are academically sound, clinical and public health laboratories seek streamlined processes without extensive checks and repeated extractions. I believe this to be an achievable goal but more work it needs to obtain desirable results from variable stool material. The comparison in this chapter was

limited to two stool samples, reflecting the exploratory nature of the study and the high resource demands of long- and short- read sequencing. While this provides useful preliminary evidence of feasibility, broader validation using additional samples will be required to establish the reproducibility and generality of these observations.

5.6 Conclusion

This study demonstrates the considerable potential of long-read metagenomic sequencing to recover complete *Campylobacter* genomes directly from stool and to deliver high-resolution strain typing, including SNP analysis and AMR profiling. When combined with high-integrity DNA extraction and adequate read length (N50 ≥ 6.8 kb), long-read data enabled contiguous genome assemblies that matched isolate-derived references in both ST and resistance determinants. However, the success of this approach is highly dependent on sequencing quality and read length, with reduced performance observed at lower N50 values. While short-read assemblies offered greater base-level accuracy and more consistent allele recovery for MLST, especially in heterogeneous samples, it is long-read sequencing that holds the unique advantage of spanning complex genomic regions and reconstructing entire genomes in a culture-free context. As sequencing chemistries and polishing tools continue to advance, long-read metagenomics is poised to become a powerful tool for direct, strain-resolved pathogen surveillance in clinical and public health settings.

6 General Discussion

Pathogen genomics has become central to public health microbiology in well-funded developed countries (Baker et al., 2023). While culture remains the gold standard for confirming the presence of infectious agents, it is often slow and can miss fastidious or unculturable organisms (Andrews & Ryan, 2015; Santos et al., 2023). The advent of WGS, particularly short-read platforms, has revolutionised outbreak detection by enabling near real-time identification, high-resolution typing, and comprehensive characterisation of pathogens, including their AMR profiles (Chattaway et al., 2019b; Neuert et al., 2018; Zhao et al., 2016). WGS offers single-nucleotide resolution that has replaced older typing methods and transformed epidemiological investigations (Joseph et al., 2023; Waldram et al., 2018). Yet, short-read approaches can struggle to assemble complete genomes (Wick et al., 2017b), resolve repetitive regions (Treangen & Salzberg, 2012), or accurately reconstruct plasmids and mobile elements (Arredondo-Alonso et al., 2017; Stadler et al., 2018), features that are often central to pathogen evolution and AMR spread. Long-read sequencing overcomes many of these limitations, providing contiguous assemblies, resolving complex genomic structures, and capturing accessory elements in a single experiment (Wick et al., 2017a). Harnessing this power in public health requires not just access to the technology but also high-quality DNA, optimised sequencing strategies, and an understanding of the trade-offs between platforms. The series of studies in this thesis address these needs by developing semi-automated HMW DNA extraction (Chapter 2), interrogating within-host variation in *Salmonella* (Chapter 3), evaluating stool preservation for metagenomics (Chapter 4), and applying long-read sequencing to recover *Campylobacter* genomes directly from human stool (Chapter 5). Together, these chapters chart a cohesive path towards more comprehensive and timely pathogen surveillance.

6.1 Methodological advancements

Long-read sequencing platforms such as ONT require high yields of long DNA fragments to maximise read length. Longer reads improve genome assembly contiguity, enable resolution of mobile elements and plasmids, and enhance detection

of antimicrobial resistance genes and structural variants that may be missed with short reads. A 96-well plate adaptation of the Fire Monkey kit was optimised on the Tecan A200 platform to yield high-molecular-weight DNA from both cultured bacteria and stool samples. This enabled high-throughput extraction of HMW DNA from clinical *Salmonella* isolates, as utilised in Chapter 3. During the project, the system was also employed to extract HMW DNA from clinical *E. coli* and *Campylobacter* isolates, demonstrating robust performance across multiple bacterial species. The HMW DNA was successfully combined with Illumina short read data to create hybrid genomes enabling analysis of structural variants. A limitation of this project was the performance of ONT's R9 chemistry for SNP analysis. The hope was to generate structurally complete and nucleotide-accurate genomes within a single hybrid assembly FASTA file, but this remained an aspirational goal. Analysis of different genome assembly strategies led me to conclude that the most accurate SNP analysis was achieved using short reads alone. ONT has since transitioned to R10 chemistry, and results from stool sequencing (Chapter 5) suggests this newer chemistry is capable of delivering high-accuracy SNP analysis as a standalone sequencing technology, and if paired with short reads will produce more reliable results than its predecessor chemistry.

The Fire Monkey protocol developed for *Salmonella* was optimised over multiple rounds and years to deliver a HMW stool DNA extraction comparable to the selected commercial benchmark (Maxwell RSC Fecal Microbiome DNA Kit). The procedures required to obtain clean DNA using the Fire Monkey kit indicate that thorough washing of the stool is essential during extraction. This was achieved by washing the stool three times in warm sterile water, increasing reagent volumes, and simultaneously reducing the input amount of stool. This protocol provides a gentle lysis approach to stool DNA extraction that avoids bead beating, offering potential for targeted pathogen long-read metagenomics in clinical and epidemiological applications. The plug-and-play design of the lysis step in this protocol allows for flexible customisation whether by optimising for specific target pathogens or incorporating enzyme cocktails to broaden extraction across bacterial species and other microbiome components such as fungi and parasites. The modifications made to the Fire Monkey protocol for stool DNA extraction render it significantly more time-consuming and labour-intensive compared to the Maxwell kit. As sample throughput increases, this could present scalability

challenges. One potential solution is the incorporation of robotic automation to perform the stool washing and lysis steps. One variable that remains unexplored due to time and funding constraints was the impact of stool input quantity (in milligrams) on the ability to detect *Campylobacter* using long-read metagenomic sequencing. During optimisation, the Fire Monkey protocol performed well in terms of yield (ng) as the stool input was gradually reduced from 200 mg to 50 mg. However, even at 50 mg, the preparation still needed to be split across multiple 2 mL tubes to complete the protocol. Reducing the input further to 10 mg would enable the entire process to be carried out in a single tube, making it far more compatible with robotic automation. Nonetheless, this raises concerns about the sensitivity of detecting *Campylobacter,* a pathogen typically present at low abundance.

6.2 Within-host genomic diversity of *Salmonella*

Standard surveillance workflows sequence a single colony per patient, implicitly assuming that the culture is clonal. This is convenient and reduces laboratory and bioinformatic workload, but it risks overlooking genetic diversity within a sample. To test this assumption, up to twenty colonies per patient were sequenced using a hybrid approach combining ONT long reads and Illumina short reads. As mentioned above, I analysed the sequencing data separately to fully leverage their respective strengths: long reads for structural resolution and short reads for single-nucleotide accuracy. Across patients, eight stool samples yielded a single *Salmonella* serovar, with six of the eight producing clonal colonies, these were colonies that differed by 1-2 SNPs. Notably, more SNP variation (5-13 SNPs) was observed in colonies from two stool samples where the serovar was *Salmonella* Java. One group of isolates showed variation in AMR determinants, with one colony having lost a set of four resistance genes due to the excision of a transposable element. Together these results show that reliance on a single colony can underestimate genomic variability and misrepresent the presence of resistance determinants. Long reads were crucial for resolving AMR genes located within repetitive regions and accurately placing them in their chromosomal context.

6.3 Impact of stool preservation on *Campylobacter* DNA

Detecting *Campylobacter* through metagenomics hinges on maintaining the integrity of both the bacterial cell structure and the DNA during storage. To evaluate the effect of preservation method, stool samples were stored under three conditions: freezing raw, freezing with broth plus glycerol, and freezing in Zymo DNA/RNA Shield. DNA was extracted pre-storage and after 1, 3, and 9 months, with recovery assessed by culture and metagenomic sequencing. Freezing raw or with glycerol generally provided the best preservation of *Campylobacter* DNA, likely reflecting reduced chemical or enzymatic degradation compared with the DNA/RNA Shield treatment. Freezing raw or with glycerol conserved overall genomic representation better, as seen by statistically significant changes in the breadth and depth of sequencing read coverage. In contrast, genome fractions from assembly did not vary significantly across conditions or timepoints. This suggests that while the method of preservation did affect how evenly and deeply the genome was sequenced, it didn't really make a big difference in the amount of the genome that could be reconstructed. This distinction highlights that read-based metrics may be more sensitive to subtle degradation effects than assembly-based measures in metagenomic recovery of *Campylobacter*.

6.4 Direct recovery of *Campylobacter* genomes from stool with long reads

In the final chapter, ONT sequencing was applied directly to DNA extracted from stool to recover *Campylobacter* genomes without prior culture. Long-read assemblies generated nearly entire genomes when *Campylobacter* DNA accounted for a significant portion of the overall metagenomic material, allowing for precise MLST typing and high-resolution SNP analysis. However, in samples where pathogen abundance was low or host DNA was dominant, the reduced proportion of *Campylobacter* reads limited assembly contiguity. This resulted in fragmented genomes and incomplete MLST profiles. Long-read platforms are incredibly powerful for bridging repetitive areas in genomic sequences, but they require adequate target coverage to accurately assemble. For such samples, deep short-read sequencing provided better base-level resolution and successfully recovered sequence types even

when long reads alone failed. These findings highlight that long-read metagenomics can enable culture-free recovery of *Campylobacter* genomes, but its effectiveness is constrained by DNA quality, host DNA background, and pathogen load. This dependency reinforces the importance of upstream factors addressed in earlier chapters, particularly optimising HMW DNA extraction and effective sample preservation, to maximise the success of long-read pathogen recovery.

6.5 Challenges facing stool metagenomics as a pathogen detection tool

Results from Chapters 4 and 5 highlight several inherent challenges to using metagenomics for pathogen surveillance in stool samples. Low pathogen abundance is a major barrier for both diagnostic sequencing and culture-based methods. Some of the limitations of can be mitigated though the use of enrichment in growth media. Issues still arise from non-viable cells and growth biases caused by additional culture steps. For metagenomic sequencing, the problem is fundamentally one of signal-to-noise: enough pathogen DNA must be present among the complex background of microbial and host DNA to achieve high-quality MDGs. In stool samples, this is further complicated by the high proportion of host DNA, which competes for sequencing depth and reduces coverage of the target organism. Pre-screening approaches, such as qPCR, can accurately predict whether a MDG is likely to be recovered, but they do not solve the underlying problem for samples with low target abundance. Emerging technologies, such as ONT's adaptive sequencing, offer potential solutions by selectively enriching target reads during sequencing. These techniques, however, rely on the availability and quality of reference databases; incomplete or poorly maintained databases run the risk of missing novel or divergent sequences, which reduces their usefulness for pathogen detection. These challenges emphasise the value of optimised upstream workflows such as those developed in earlier chapters for DNA extraction and preservation to maximise pathogen signal before sequencing begins.

6.6 Public health implications and future outlook

A framework for improving pathogen surveillance is provided by this study, which integrates methodological advancements from other chapters. Automated HMW DNA extraction enables scalable preparation for both long-read sequencing of isolates and metagenomic analysis of stool. Hybrid sequencing strategies reveal within-host diversity and deliver accurate genome characterisation, demonstrating that single-colony approaches can overlook clinically relevant variation. Sample handling procedures are informed by preservation studies; for metagenomics, quick freezing in glycerol is advised to preserve DNA integrity, while chemical stabilisers like DNA/RNA Shield may make it more difficult to detect pathogens. Direct long-read ONT sequencing from stool samples shows great potential for diagnosing fastidious organisms like *Campylobacter* without the need for culture. This is, of course, contingent on having good DNA quality and a sufficient amount of the pathogen present. As sequencing technologies continue to evolve, the findings presented here, particularly the critical role of high-integrity input DNA and the complementary strengths of long- and short-read platforms, will remain essential for delivering accurate, timely, and actionable genomic data in public health contexts.

Future progress will depend on integrating these approaches into routine laboratory workflows, improving reference database curation to support adaptive sequencing and accurate typing, and developing cost-effective protocols for low-abundance pathogens. Advances in real-time analysis pipelines, combined with robust sample extraction methods, could enable near-instant genomic surveillance at the point of care. Ultimately, bridging these technical and operational gaps will be key to translating metagenomic potential into a reliable frontline tool for global public health.

6.7 Future directions

6.7.1 General

Building on the findings of this work, several avenues for future research could further strengthen the role of metagenomics in pathogen surveillance. Other cryoprotectants and stabilisers that preserve DNA integrity without obstructing downstream detection could be investigated in order to enhance preservation techniques. This would not only be of benefit for *Campylobacter* but also for a wider variety of pathogens and low-abundance species. Parallel efforts are needed to improve host DNA depletion in stool metagenomics, testing physical, enzymatic, and adaptive sequencing-based methods to increase the proportion of pathogen reads in low-load samples. Enhancing long-read metagenomics for such low-abundance targets will require refining library preparation protocols for minimal input DNA while maintaining read length and quality and developing targeted enrichment strategies powered by well-curated reference databases.

At the workflow level, integrating hybrid long- and short-read sequencing into public health laboratories offers the potential for comprehensive and accurate genomic surveillance, provided that cost, turnaround time, and automated data processing pipelines are optimised for routine use. The combination of rapid extraction, host depletion, and portable long-read sequencing could also enable real-time genomic surveillance at the point of care, particularly in outbreak or low-resource settings. Finally, future studies should explore how genomic data, whether derived from metagenomics or culture, can be more effectively linked to clinical and epidemiological information, ensuring that improvements in laboratory capability translate into faster, more informed public health responses.

6.7.2 Focused

A finding in Chapter 3 suggested that specific *Salmonella* sequence types may have disappeared from circulation in the UK during the COVID-19 pandemic, potentially driven by unprecedented global changes to human behaviour, travel, and food supply chains. Extending this analysis across all *Salmonella* serovars, using the comprehensive genomic datasets in EnteroBase, could yield unique insights into how large-scale societal disruptions reshape pathogen populations. Such an investigation could systematically document sequence types that remained in constant circulation,

those that disappeared entirely, those that disappeared and later re-emerged, and those that increased in prevalence or became dominant during this period.

Two key opportunities emerge from the *Salmonella* within-patient diversity work. First, repeating the analysis with ONT R10 chemistry would test whether the latest long-read technology can independently deliver the resolution needed for robust SNP-based analyses, and whether, when paired with short reads, it can produce hybrid assemblies of sufficient quality for high-confidence genomic epidemiology. Second, a large-scale investigation of genome-level variation in patients with salmonellosis covering a wide diversity of serovars could reveal patterns of within-host evolution and identify genes under selective pressure, providing valuable insights into pathogen adaptation and persistence during infection.

Unfortunately, the combination of the Fire Monkey platform and the Tecan A200 will not be made available for routine use. From a Quadram perspective, however, several lessons learned during its development particularly regarding stool washing and pre-processing could be adapted and tested in combination with the Promega Maxwell systems. Applying these optimisations to the Maxwell workflow may improve inhibitor removal, enhance DNA yield and integrity, and ultimately increase the success rate of downstream sequencing, particularly for challenging metagenomic samples.

6.8 Final remarks

Overall, this work has significantly advanced the understanding of how methodological choices in DNA extraction, sample preservation, and sequencing strategy shape the recovery and resolution of pathogen genomes from complex clinical samples. By developing and optimising workflows for high-molecular-weight DNA extraction, evaluating preservation methods for metagenomics, and applying both long- and short-read sequencing to real-world public health challenges, it has provided practical, evidence-based guidance for improving pathogen surveillance. The integration of hybrid sequencing approaches, insights into within-host diversity, and demonstration of culture-independent genome recovery from stool collectively offer a roadmap for more genomic epidemiology. These findings not only strengthen the technical foundations of pathogen genomics but also open new avenues for

epidemiological investigation, ensuring that future advances in sequencing technologies can be effectively translated into actionable public health impact.

7 References

Achtman, M., Wain, J., Weill, F.-X., Nair, S., Zhou, Z., Sangal, V., Krauland, M. G., Hale, J. L., Harbottle, H., & Uesbeck, A. (2012). Multilocus sequence typing as a replacement for serotyping in *Salmonella enterica*. *PLoS pathogens*, *8*(6), e1002776.

Adeleye, S. A., & Yadavalli, S. S. (2024). Queuosine biosynthetic enzyme, QueE moonlights as a cell division regulator. *PLoS genetics*, *20*(5), e1011287.

Ahlinder, J., Svedberg, A. L., Nystedt, A., Dryselius, R., Jacobsson, K., Hägglund, M., Brindefalk, B., Forsman, M., Ottoson, J., & Troell, K. (2022). Use of metagenomic microbial source tracking to investigate the source of a foodborne outbreak of cryptosporidiosis. *Food and Waterborne Parasitology*, *26*. https://doi.org/ARTN e00142
10.1016/j.fawpar.2021.e00142

Alikhan, N.-F., Zhou, Z., Sergeant, M. J., & Achtman, M. (2018). A genomic overview of the population structure of *Salmonella*. *PLoS genetics*, *14*(4), e1007261.

Allard, M. W. (2016). The Future of Whole-Genome Sequencing for Public Health and the Clinic. *Journal of Clinical Microbiology*, *54*(8), 1946-1948. https://doi.org/10.1128/Jcm.01082-16

Andersen, S. C., & Hoorfar, J. (2018). Surveillance of Foodborne Pathogens: Towards Diagnostic Metagenomics of Fecal Samples. *Genes*, *9*(1). https://doi.org/ARTN 14
10.3390/genes9010014

Andrews, J. R., & Ryan, E. T. (2015). Diagnostics for invasive *Salmonella* infections: Current challenges and future directions. *Vaccine*, *33*, C8-C15. https://doi.org/10.1016/j.vaccine.2015.02.030

Anjum, M. F. (2015). Screening methods for the detection of antimicrobial resistance genes present in bacterial isolates and the microbiota. *Future microbiology*, *10*(3), 317-320.

Anthony, C., Pearson, K., Callaby, R., Allison, L., Jenkins, C., Smith-Palmer, A., & James, M. (2024). Reasons for difficulties in isolating the causative organism during food-borne outbreak investigations using STEC as a model pathogen: a systematic review, 2000 to 2019. *Eurosurveillance*, *29*(49). https://doi.org/Artn 2400193
10.2807/1560-7917.Es.2024.29.49.2400193

Antipov, D., Raiko, M., Lapidus, A., & Pevzner, P. A. (2019). Plasmid detection and assembly in genomic and metagenomic data sets. *Genome Research*, *29*(6), 961-968. https://doi.org/10.1101/gr.241299.118

Arning, N., Sheppard, S. K., Bayliss, S., Clifton, D. A., & Wilson, D. J. (2021). Machine learning to predict the source of campylobacteriosis using whole genome data. *PLoS genetics*, *17*(10). https://doi.org/ARTN e1009436
10.1371/journal.pgen.1009436

Arredondo-Alonso, S., Willems, R. J., van Schaik, W., & Schürch, A. C. (2017). On the (im)possibility of reconstructing plasmids from whole-genome short-read sequencing data. *Microbial Genomics*, *3*(10). https://doi.org/ARTN 000128
10.1099/mgen.0.000128
Auguet, O. T., Niehus, R., Gweon, H. S., Berkley, J. A., Waichungo, J., Njim, T., Edgeworth, J. D., Batra, R., Chau, K., Swann, J., Walker, S. A., Peto, T. E. A., Crook, D. W., Lamble, S., Turner, P., Cooper, B. S., & Stoesser, N. (2021). Population-level faecal metagenomic profiling as a tool to predict antimicrobial resistance in Enterobacterales isolates causing invasive infections: An exploratory study across Cambodia, Kenya, and the UK. *Eclinicalmedicine*, *36*. https://doi.org/ARTN 100910
10.1016/j.eclinm.2021.100910
Authority, E. F. S. (2021). Multi-country outbreak of multiple *Salmonella enterica* serotypes linked to imported sesame-based products. *EFSA Supporting Publications*, *18*(10), 6922E.
Baker, K. S., Jauneikaite, E., Hopkins, K. L., Lo, S. W., Sánchez-Busór, L., Getino, M., Howden, B. P., Holt, K. E., Musila, L. A., Hendriksen, R. S., Amoako, D. G., Aanensen, D. M., Okeke, I. N., Egyir, B., Nunn, J. G., Midega, J. T., Feasey, N. A., Peacock, S. J., & Surveillance, S. G. (2023). Genomics for public health and international surveillance of antimicrobial resistance. *Lancet Microbe*, *4*(12), e1047-e1055. https://doi.org/10.1016/S2666-5247(23)00283-5
Balmer, O., & Tanner, M. (2011). Prevalence and implications of multiple-strain infections. *The Lancet infectious diseases*, *11*(11), 868-878.
Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., & Prjibelski, A. D. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of computational biology*, *19*(5), 455-477.
Barbosa, C., Nogueira, S., Gadanho, M., & Chaves, S. (2016). DNA extraction: finding the most suitable method. In *Molecular microbial diagnostic methods* (pp. 135-154). Elsevier.
Barker, C. R., Painset, A., Swift, C., Jenkins, C., Godbole, G., Maiden, M. C. J., & Dallman, T. J. (2020). Microevolution of *Campylobacter jejuni* during long-term infection in an immunocompromised host. *Scientific reports*, *10*(1). https://doi.org/ARTN 10109
10.1038/s41598-020-66771-7
Bastidas-Caldes, C., de Waard, J. H., Salgado, M. S., Villacís, M. J., Coral-Almeida, M., Yamamoto, Y., & Calvopiña, M. (2022). Worldwide prevalence of mcr-mediated colistin-resistance *Escherichia coli* in isolates of clinical samples, healthy humans, and livestock—a systematic review and meta-analysis. *Pathogens*, *11*(6), 659.
Batool, M., & Galloway-Peña, J. (2023). Clinical metagenomics—challenges and future prospects. *Frontiers in Microbiology*, *14*, 1186424.
Becker, L., Steglich, M., Fuchs, S., Werner, G., & Nübel, U. (2016). Comparison of six commercial kits to extract bacterial chromosome and plasmid

DNA for MiSeq sequencing. *Scientific reports*, *6*. https://doi.org/ARTN 28063
10.1038/srep28063
Benoit, G., Raguideau, S., James, R., Phillippy, A. M., Chikhi, R., & Quince, C. (2024). High-quality metagenome assembly from long accurate reads with metaMDBG. *Nature Biotechnology*, *42*(9), 1378-1383.
Benoit, P., Brazer, N., de Lorenzi-Tognon, M., Kelly, E., Servellita, V., Oseguera, M., Nguyen, J., Tang, J. C., Omura, C., Streithorst, J., Hillberg, M., Ingebrigtsen, D., Zorn, K., Wilson, M. R., Blicharz, T., Wong, A. P., O'Donovan, B., Murray, B., Miller, S., & Chiu, C. Y. (2024). Seven-year performance of a clinical metagenomic next-generation sequencing test for diagnosis of central nervous system infections. *Nature Medicine*, *30*(12). https://doi.org/10.1038/s41591-024-03275-1
Berbers, B., Saltykova, A., Garcia-Graells, C., Philipp, P., Arella, F., Marchal, K., Winand, R., Vanneste, K., Roosens, N. H. C., & De Keersmaecker, S. C. J. (2020). Combining short and long read sequencing to characterize antimicrobial resistance genes on plasmids applied to an unauthorized genetically modified. *Scientific reports*, *10*(1). https://doi.org/ARTN 4310
10.1038/s41598-020-61158-0
Berbers, B., Vanneste, K., Roosens, N. H. C. J., Marchal, K., Ceyssens, P. J., & De Keersmaecker, S. C. J. (2023). Using a combination of short- and long-read sequencing to investigate the diversity in plasmid- and chromosomally encoded extended-spectrum beta-lactamases (ESBLs) in clinical *Shigella* and *Salmonella* isolates in Belgium. *Microbial Genomics*, *9*(1). https://doi.org/10.1099/mgen.0.000925
Bertrand, D., Shaw, J., Kalathiyappan, M., Ng, A. H. Q., Kumar, M. S., Li, C., Dvornicic, M., Soldo, J. P., Koh, J. Y., & Tong, C. (2019). Hybrid metagenomic assembly enables high-resolution analysis of resistance determinants and mobile elements in human microbiomes. *Nature Biotechnology*, *37*(8), 937-944.
Beutin, L., & Martin, A. (2012). Outbreak of Shiga Toxin–Producing *Escherichia coli* (STEC) O104:H4 Infection in Germany Causes a Paradigm Shift with Regard to Human Pathogenicity of STEC Strains. *Journal of Food Protection*, *75*(2), 408-418. https://doi.org/10.4315/0362-028x.Jfp-11-452
Black, A., MacCannell, D. R., Sibley, T. R., & Bedford, T. (2020). Ten recommendations for supporting open pathogen genomic analysis in public health. *Nature Medicine*, *26*(6), 832-841. https://doi.org/10.1038/s41591-020-0935-z
Blanco-Míguez, A., Beghini, F., Cumbo, F., McIver, L. J., Thompson, K. N., Zolfo, M., Manghi, P., Dubois, L., Huang, K. D., Thomas, A. M., Nickols, W. A., Piccinno, G., Piperni, E., Puncochár, M., Valles-Colomer, M., Tett, A., Giordano, F., Davies, R., Wolf, J., . . . Segata, N. (2023). Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn 4. *Nature Biotechnology*, *41*(11), 1633-+. https://doi.org/ARTN s41587-023-01688-w
10.1038/s41587-023-01688-w

Bloomfield, S. J., Zomer, A. L., O'grady, J., Kay, G. L., Wain, J., Janecko, N., Palau, R., & Mather, A. E. (2023). Determination and quantification of microbial communities and antimicrobial resistance on food through host DNA-depleted metagenomics. *Food Microbiology*, *110*, 104162.

Bogaerts, B., Van den Bossche, A., Verhaegen, B., Delbrassinne, L., Mattheus, W., Nouws, S., Godfroid, M., Hoffman, S., Roosens, N. H., & De Keersmaecker, S. C. (2024). Closing the gap: Oxford Nanopore Technologies R10 sequencing allows comparable results to Illumina sequencing for SNP-based outbreak investigation of bacterial pathogens. *Journal of Clinical Microbiology*, *62*(5), e01576-01523.

Bouras, G., Houtak, G., Wick, R. R., Mallawaarachchi, V., Roach, M. J., Papudeshi, B., Judd, L. M., Sheppard, A. E., Edwards, R. A., & Vreugde, S. (2024). Hybracter: enabling scalable, automated, complete and accurate bacterial genome assemblies. *Microbial Genomics*, *10*(5), 001244.

Bronner, I. F., Dawson, E., Park, N., Piepenburg, O., & Quail, M. A. (2025). Evaluation of controls, quality control assays, and protocol optimisations for PacBio HiFi sequencing on diverse and challenging samples. *Frontiers in Genetics*, *15*, 1505839.

Brown, E., Dessai, U., McGarry, S., & Gerner-Smidt, P. (2019). Use of whole-genome sequencing for food safety and public health in the United States. *Foodborne pathogens and disease*, *16*(7), 441-450.

Bukari, Z., Emmanuel, T., Woodward, J., Ferguson, R., Ezughara, M., Darga, N., & Lopes, B. S. (2025). The Global Challenge of *Campylobacter*: Antimicrobial Resistance and Emerging Intervention Strategies. *Tropical Medicine and Infectious Disease*, *10*(1), 25.

Cao, M. D., Nguyen, S. H., Ganesamoorthy, D., Elliott, A. G., Cooper, M. A., & Coin, L. J. (2017). Scaffolding and completing genome assemblies in real-time with nanopore sequencing. *Nature Communications*, *8*, 14515. https://doi.org/10.1038/ncomms14515

Cardona, S., Eck, A., Cassellas, M., Gallart, M., Alastrue, C., Dore, J., Azpiroz, F., Roca, J., Guarner, F., & Manichanh, C. (2012). Storage conditions of intestinal microbiota matter in metagenomic analysis. *Bmc Microbiology*, *12*. https://doi.org/Artn 158
10.1186/1471-2180-12-158

Carter, C., Hutchison, A., Rudder, S., Trotter, E., Waters, E. V., Elumogo, N., & Langridge, G. C. (2023). Uropathogenic *Escherichia coli* population structure and antimicrobial susceptibility in Norfolk, UK. *Journal of antimicrobial chemotherapy*, *78*(8), 2028-2036.

Carver, T., Berriman, M., Tivey, A., Patel, C., Böhme, U., Barrell, B. G., Parkhill, J., & Rajandream, M.-A. (2008). Artemis and ACT: viewing, annotating and comparing sequences stored in a relational database. *Bioinformatics*, *24*(23), 2672-2676.

Cayrou, C., Barratt, N. A., Ketley, J. M., & Bayliss, C. D. (2021). Phase Variation During Host Colonization and Invasion by Campylobacter jejuni and Other Campylobacter Species. *Front Microbiol*, *12*, 705139. https://doi.org/10.3389/fmicb.2021.705139

Chachaty, E., & Saulnier, P. (2000). Isolating chromosomal DNA from bacteria. *The nucleic acid protocols handbook*, 29-32.

Charalampous, T., Kay, G. L., Richardson, H., Aydin, A., Baldan, R., Jeanes, C., Rae, D., Grundy, S., Turner, D. J., & Wain, J. (2019). Nanopore metagenomics enables rapid clinical diagnosis of bacterial lower respiratory infection. *Nature biotechnology*, *37*(7), 783-792.

Charalampous, T., Kay, G. L., Richardson, H., Aydin, A., Baldan, R., Jeanes, C., Rae, D., Grundy, S., Turner, D. J., Wain, J., Leggett, R. M., Livermore, D. M., & O'Grady, J. (2019). Nanopore metagenomics enables rapid clinical diagnosis of bacterial lower respiratory infection. *Nature Biotechnology*, *37*(7), 783-+. https://doi.org/10.1038/s41587-019-0156-5

Chattaway, M. A., Dallman, T. J., Larkin, L., Nair, S., McCormick, J., Mikhail, A., Hartman, H., Godbole, G., Powell, D., Day, M., Smith, R., & Grant, K. (2019a). The Transformation of Reference Microbiology Methods and Surveillance for Salmonella With the Use of Whole Genome Sequencing in England and Wales. *Frontiers in public health*, *7*, 317.

Chattaway, M. A., Dallman, T. J., Larkin, L., Nair, S., McCormick, J., Mikhail, A., Hartman, H., Godbole, G., Powell, D., Day, M., Smith, R., & Grant, K. (2019b). The Transformation of Reference Microbiology Methods and Surveillance for *Salmonella* With the Use of Whole Genome Sequencing in England and Wales. *Frontiers in public health*, *7*. https://doi.org/ARTN 317
10.3389/fpubh.2019.00317

Chattaway, M. A., Painset, A., Godbole, G., Gharbia, S., & Jenkins, C. (2023). Evaluation of genomic typing methods in the *Salmonella* reference laboratory in Public Health, England, 2012–2020. *Pathogens*, *12*(2), 223.

Chaumeil, P. A., Mussig, A. J., Hugenholtz, P., & Parks, D. H. (2019). GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics*, *36*(6), 1925-1927. https://doi.org/10.1093/bioinformatics/btz848

Chen, A. L., Hu, Y. X., Zhang, Y. J., Li, Z. J., Zeng, Y., & Pang, X. Y. (2022). Cryopreservation of stool samples altered the microbial viability quantitively and compositionally. *Archives of Microbiology*, *204*(9). https://doi.org/ARTN 557
10.1007/s00203-022-03169-1

Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, *34*(17), i884-i890.

Chen, Y., Liu, Y., Shi, Y., Ping, J., Wu, J., & Chen, H. (2020). Magnetic particles for integrated nucleic acid purification, amplification and detection without pipetting. *TrAC Trends in Analytical Chemistry*, *127*, 115912.

Chen, Z., Erickson, D. L., & Meng, J. (2021). Polishing the Oxford Nanopore long-read assemblies of bacterial pathogens with Illumina short reads to improve genomic analyses. *Genomics*, *113*(3), 1366-1377.

Chiang, A. D., & Palmore, T. N. (2022). Whole Genome Sequencing for Outbreak Investigation. In *Infection Prevention: New Perspectives and Controversies* (pp. 223-235). Springer.

Chiu, C. Y., & Miller, S. A. (2019). Clinical metagenomics. *Nature reviews genetics*, *20*(6), 341-355. https://doi.org/10.1038/s41576-019-0113-7

Chlebicz, A., & Slizewska, K. (2018). Campylobacteriosis, Salmonellosis, Yersiniosis, and Listeriosis as Zoonotic Foodborne Diseases: A Review. *International Journal of Environmental Research and Public Health*, *15*(5). https://doi.org/ARTN 863
10.3390/ijerph15050863

Choo, J. M., Leong, L. E. X., & Rogers, G. B. (2015). Sample storage conditions significantly influence faecal microbiome profiles. *Scientific reports*, *5*. https://doi.org/ARTN 16350
10.1038/srep16350

Cody, A. J., McCarthy, N. D., Jansen van Rensburg, M., Isinkaye, T., Bentley, S. D., Parkhill, J., Dingle, K. E., Bowler, I. C., Jolley, K. A., & Maiden, M. C. (2013). Real-time genomic epidemiological evaluation of human *Campylobacter* isolates by use of whole-genome multilocus sequence typing. *Journal of Clinical Microbiology*, *51*(8), 2526-2534.

Cooper, A. L., Low, A. J., Koziol, A. G., Thomas, M. C., Leclair, D., Tamber, S., Wong, A., Blais, B. W., & Carrillo, C. D. (2020). Systematic Evaluation of Whole Genome Sequence-Based Predictions of *Salmonella* Serotype and Antimicrobial Resistance. *Frontiers in Microbiology*, *11*. https://doi.org/ARTN 549
10.3389/fmicb.2020.00549

Costa, D., & Iraola, G. (2019). Pathogenomics of Emerging *Campylobacter* Species. *Clinical Microbiology Reviews*, *32*(4). https://doi.org/ARTN e00072-18
10.1128/CMR.00072-18

Costantini, V., Grenz, L., Fritzinger, A., Lewis, D., Biggs, C., Hale, A., & Vinjé, J. (2010). Diagnostic Accuracy and Analytical Sensitivity of IDEIA Norovirus Assay for Routine Screening of Human Norovirus. *Journal of Clinical Microbiology*, *48*(8), 2770-2778. https://doi.org/10.1128/Jcm.00654-10

Cruz, E., Haeberle, A., Westerman, T., Durham, M., Suyemoto, M., Knodler, L., & Elfenbein, J. (2023). Nonredundant dimethyl sulfoxide reductases influence *Salmonella enterica* serotype Typhimurium anaerobic growth and virulence. *Infection and immunity*, *91*(2), e00578-00522.

Curiao, T., Marchi, E., Grandgirard, D., León-Sampedro, R., Viti, C., Leib, S. L., Baquero, F., Oggioni, M. R., Martinez, J. L., & Coque, T. M. (2016). Multiple adaptive routes of *Salmonella enterica* Typhimurium to biocide and antibiotic exposure. *Bmc Genomics*, *17*, 1-16.

Czibener, C., Merwaiss, F., Guaimas, F., Del Giudice, M. G., Serantes, D. A. R., Spera, J. M., & Ugalde, J. E. (2016). BigA is a novel adhesin of *Brucella* that mediates adhesion to epithelial cells. *Cellular microbiology*, *18*(4), 500-513.

Danaeifar, M. (2022). New horizons in developing cell lysis methods: A review. *Biotechnology and Bioengineering*, *119*(11), 3007-3021. https://doi.org/10.1002/bit.28198

Dasti, J. I., Gross, U., Pohl, S., Lugert, R., Weig, M., & Schmidt-Ott, R. (2007). Role of the plasmid-encoded (O) gene in tetracycline-resistant clinical

isolates of and *Campylobacter jejuni* and *Campylobacter coli*. *Journal of Medical Microbiology*, *56*(6), 833-837. https://doi.org/10.1099/jmm.0.47103-0

Davati, N., Ghorbani, A., Ashrafi-Dehkordi, E., & Karbanowicz, T. P. (2023). Gene networks analysis of *Salmonella* Typhimurium reveals new insights on key genes involved in response to low water activity. *Iranian Journal of Biotechnology*, *21*(4), e3640.

Daza-Cardona, E. A., Buenhombre, J., dos Santos Fontenelle, R. O., & Barbosa, F. C. B. (2022). *mcr*-mediated colistin resistance in South America, a One Health approach: a review. *Reviews and Research in Medical Microbiology*, *33*(1), e119-e136.

de Bruin, O. M., Chiefari, A., Wroblewski, D., Egan, C., & Kelly-Cirino, C. D. (2019). A novel chemical lysis method for maximum release of DNA from difficult-to-lyse bacteria. *Microbial Pathogenesis*, *126*, 292-297. https://doi.org/10.1016/j.micpath.2018.11.008

De Coster, W., D'Hert, S., Schultz, D. T., Cruts, M., & Van Broeckhoven, C. (2018). NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics*, *34*(15), 2666-2669. https://doi.org/10.1093/bioinformatics/bty149

de la Gandara, P. (2023). Countrywide multi-serotype outbreak of *Salmonella* Bovismorbificans ST142 and monophasic *Salmonella* Typhimurium ST34 associated with dried pork sausages in France, September to January 2021. *Eurosurveillance*, *28*(3). <Go to ISI>://WOS:001045151700001

De, R. (2019). Metagenomics: aid to combat antimicrobial resistance in diarrhea. *Gut Pathogens*, *11*(1). https://doi.org/ARTN 47 10.1186/s13099-019-0331-8

Deng, X., den Bakker, H. C., & Hendriksen, R. S. (2016). Genomic epidemiology: whole-genome-sequencing–powered surveillance and outbreak investigation of foodborne bacterial pathogens. *Annual review of food science and technology*, *7*(1), 353-374.

Destoumieux-Garzón, D., Mavingui, P., Boetsch, G., Boissier, J., Darriet, F., Duboz, P., Fritsch, C., Giraudoux, P., Le Roux, F., Morand, S., Paillard, C., Pontier, D., Sueur, C., & Voituron, Y. (2018). The One Health Concept: 10 Years Old and a Long Road Ahead. *Frontiers in Veterinary Science*, *5*. https://doi.org/ARTN 14 10.3389/fvets.2018.00014

Dicksved, J., Ellström, P., Engstrand, L., & Rautelin, H. (2014). Susceptibility to Infection Is Associated with the Species Composition of the Human Fecal Microbiota. *MBio*, *5*(5). https://doi.org/ARTN e01212-14 10.1128/mBio.01212-14

Djeghout, B., Bloomfield, S. J., Rudder, S., Elumogo, N., Mather, A. E., Wain, J., & Janecko, N. (2022). Comparative genomics of *Campylobacter jejuni* from clinical campylobacteriosis stool specimens. *Gut Pathogens*, *14*(1), 45.

Djeghout, B., Le-Viet, T., Martins, L. D., Savva, G. M., Evans, R., Baker, D., Page, A., Elumogo, N., Wain, J., & Janecko, N. (2024). Capturing clinically relevant *Campylobacter* attributes through direct whole genome

sequencing of stool. *Microbial Genomics*, *10*(8). https://doi.org/ARTN 001284
10.1099/mgen.0.001284
Djeghout, B., Le-Viet, T., Martins, L. d. O., Savva, G. M., Evans, R., Baker, D., Page, A., Elumogo, N., Wain, J., & Janecko, N. (2024). Capturing clinically relevant *Campylobacter* attributes through direct whole genome sequencing of stool. *Microbial Genomics*, *10*(8), 001284.
Djeghout, B., Rudder, S. J., Le Viet, T., Elumogo, N., Langridge, G., & Janecko, N. (2025). Case study: Genomic characteristics of the gut microbiome, *Campylobacter* and *Salmonella* genotypes in three cases of gastroenteritis co-infections. *bioRxiv*, 2025.2004. 2029.651233.
Dore, J., Ehrlich, S., Levenez, F., Pelletier, E., Alberti, A., Bertrand, L., Bork, P., Costea, P., Sunagawa, S., & Guarner, F. (2015). Standard Operating Procedure for Fecal Samples DNA Extraction. *Protocol Q. International Human Microbiome Standards*.
Dupuy, E., & Collet, J.-F. (2021). Fort CnoX: Protecting bacterial proteins from misfolding and oxidative damage. *Frontiers in molecular biosciences*, *8*, 681932.
Durack, J., Burke, T. P., & Portnoy, D. A. (2015). A prl mutation in SecY suppresses secretion and virulence defects of *Listeria monocytogenes* secA2 mutants. *Journal of bacteriology*, *197*(5), 932-942.
Dziegiel, A. H., Bloomfield, S. J., Savva, G. M., Palau, R., Janecko, N., Wain, J., & Mather, A. E. (2024). High *Campylobacter* diversity in retail chicken: epidemiologically important strains may be missed with current sampling methods. *Epidemiology & Infection*, *152*, e101.
Eagle, S. H., Robertson, J., Bastedo, D. P., Liu, K., & Nash, J. H. (2023). Evaluation of five commercial DNA extraction kits using *Salmonella* as a model for implementation of rapid Nanopore sequencing in routine diagnostic laboratories. *Access Microbiology*, *5*(2), 000468. v000463.
Elek, C. K. A., Brown, T. L., Le Viet, T., Evans, R., Baker, D. J., Telatin, A., Tiwari, S. K., Al-Khanaq, H., Thilliez, G., Kingsley, R. A., Hall, L. J., Webber, M. A., & Adriaenssens, E. M. (2023). A hybrid and polypolish workflow for the complete and accurate assembly of phage genomes: a case study of ten przondoviruses. *Microbial Genomics*, *9*(7). https://doi.org/ARTN 001065
10.1099/mgen.0.001065
Espinosa, E., Bautista, R., Larrosa, R., & Plata, O. (2024). Advancements in long-read genome sequencing technologies and algorithms. *Genomics*, 110842.
Eyre, D. W., Cule, M. L., Griffiths, D., Crook, D. W., Peto, T. E., Walker, A. S., & Wilson, D. J. (2013). Detection of mixed infection from bacterial whole genome sequence data allows assessment of its role in *Clostridium difficile* transmission. *Plos Computational Biology*, *9*(5), e1003059.
Facciolà, A., Riso, R., Avventuroso, E., Visalli, G., Delia, S. A., & Lagana, P. (2017). *Campylobacter*: from microbiology to prevention. *Journal of preventive medicine and hygiene*, *58*(2), E79.
Fernández-Pato, A., Sinha, T., Gacesa, R., Andreu-Sánchez, S., Gois, M. F. B., Gelderloos-Arends, J., Jansen, D. B. H., Kruk, M., Jaeger, M., Joosten, L.

A. B., Netea, M. G., Weersma, R. K., Wijmenga, C., Harmsen, H. J. M., Fu, J. Y., Zhernakova, A., & Kurilshikov, A. (2024). Choice of DNA extraction method affects stool microbiome recovery and subsequent phenotypic association analyses. *Scientific reports*, *14*(1). https://doi.org/ARTN 3911
10.1038/s41598-024-54353-w

Fiedorová, K., Radvansky, M., Nemcová, E., Grombiríková, H., Bosák, J., Cernochová, M., Lexa, M., Smajs, D., & Freiberger, T. (2019). The Impact of DNA Extraction Methods on Stool Bacterial and Fungal Microbiota Community Recovery. *Frontiers in Microbiology*, *10*. https://doi.org/ARTN 821
10.3389/fmicb.2019.00821

Fitzgerald, C., Patrick, M., Gonzalez, A., Akin, J., Polage, C. R., Wymore, K., Gillim-Ross, L., Xavier, K., Sadlowski, J., Monahan, J., Hurd, S., Dahlberg, S., Jerris, R., Watson, R., Santovenia, M., Mitchell, D., Harrison, C., Tobin-D'Angelo, M., DeMartino, M., . . . Study, C. D. (2016). Multicenter evaluation of clinical diagnostic methods for detection and isolation of *Campylobacter* spp. from stool. *Journal of Clinical Microbiology*, *54*(5), 1209-1215. https://doi.org/10.1128/Jcm.01925-15

Forbes, J. D., Knox, N. C., Ronholm, J., Pagotto, F., & Reimer, A. (2017). Metagenomics: The Next Culture-Independent Game Changer. *Frontiers in Microbiology*, *8*. https://doi.org/ARTN 1069
10.3389/fmicb.2017.01069

Ford, L., Carter, G. P., Wang, Q., Seemann, T., Sintchenko, V., Glass, K., Williamson, D. A., Howard, P., Valcanis, M., & Castillo, C. F. S. (2018). Incorporating whole-genome sequencing into public health surveillance: lessons from prospective sequencing of *Salmonella* Typhimurium in Australia. *Foodborne pathogens and disease*, *15*(3), 161-167.

Franklin, N., Hope, K., Glasgow, K., & Glass, K. (2020). Describing the Epidemiology of Foodborne Outbreaks in New South Wales from 2000 to 2017. *Foodborne pathogens and disease*, *17*(11), 701-711. https://doi.org/10.1089/fpd.2020.2806

Furuse, Y. (2021). Genomic sequencing effort for SARS-CoV-2 by country during the pandemic. *International Journal of Infectious Diseases*, *103*, 305-307. https://doi.org/10.1016/j.ijid.2020.12.034

Galán-Relaño, Á., Valero Díaz, A., Huerta Lorenzo, B., Gómez-Gascón, L., Mena Rodríguez, M. Á., Carrasco Jiménez, E., Pérez Rodríguez, F., & Astorga Márquez, R. J. (2023). *Salmonella* and salmonellosis: An update on public health implications and control strategies. *Animals*, *13*(23), 3666.

Gand, M., Bloemen, B., Vanneste, K., Roosens, N. H. C., & De Keersmaecker, S. C. J. (2023). Comparison of 6 DNA extraction methods for isolation of high yield of high molecular weight DNA suitable for shotgun metagenomics Nanopore sequencing to detect bacteria. *Bmc Genomics*, *24*(1), 438.

Gautam, A. (2022). DNA isolation by lysozyme and proteinase K. In *DNA and RNA Isolation Techniques for Non-Experts* (pp. 85-88). Springer.

Gehrig, J. L., Portik, D. M., Driscoll, M. D., Jackson, E., Chakraborty, S., Gratalo, D., Ashby, M., & Valladares, R. (2022). Finding the right fit: evaluation of short-read and long-read sequencing approaches to maximize the utility of clinical microbiome data. *Microbial Genomics*, *8*(3), 000794.

Gilchrist, C. A., Turner, S. D., Riley, M. F., Petri Jr, W. A., & Hewlett, E. L. (2015). Whole-genome sequencing in outbreak analysis. *Clinical Microbiology Reviews*, *28*(3), 541-563.

Golz, J. C., & Stingl, K. (2021). Natural competence and horizontal gene transfer in *Campylobacter*. *Fighting Campylobacter Infections: Towards a One Health Approach*, 265-292.

Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nature reviews genetics*, *17*(6), 333-351.

Gorman, R., & Adley, C. C. (2004). An evaluation of five preservation techniques and conventional freezing temperatures of -20°C and -85°C for long-term preservation of. *Letters in Applied Microbiology*, *38*(4), 306-310. https://doi.org/10.1111/j.1472-765X.2004.01490.x

Grad, Y. H., Lipsitch, M., Feldgarden, M., Arachchi, H. M., Cerqueira, G. C., FitzGerald, M., Godfrey, P., Haas, B. J., Murphy, C. I., & Russ, C. (2012). Genomic epidemiology of the *Escherichia coli* O104: H4 outbreaks in Europe, 2011. *Proceedings of the national academy of sciences*, *109*(8), 3065-3070.

Granja-Salcedo, Y. T., Ramirez-Uscategui, R. A., Machado, E. G., Messana, J. D., Kishi, L. T., Dias, A. V. L., & Berchielli, T. T. (2017). Studies on bacterial community composition are affected by the time and storage method of the rumen content. *Plos One*, *12*(4). https://doi.org/ARTN e0176701

10.1371/journal.pone.0176701

Greig, D. R., Bird, M. T., Chattaway, M. A., Langridge, G. C., Waters, E. V., Ribeca, P., Jenkins, C., & Nair, S. (2022). Characterization of a P1-bacteriophage-like plasmid (phage-plasmid) harbouring bla CTX-M-15 in *Salmonella enterica* serovar Typhi. *Microbial Genomics*, *8*(12), 000913.

Guerin-Danan, C. (1999). Storage of intestinal bacteria in samples frozen with glycerol. *Microbial Ecology in Health and Disease*, *11*(3), 180-182.

Guo, A., Cao, S., Tu, L., Chen, P., Zhang, C., Jia, A., Yang, W., Liu, Z., Chen, H., & Schifferli, D. M. (2009). FimH alleles direct preferential binding of *Salmonella* to distinct mammalian cells or to avian cells. *Microbiology*, *155*(5), 1623-1633.

Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. *Bioinformatics*, *29*(8), 1072-1075. https://doi.org/10.1093/bioinformatics/btt086

Hadfield, J., Megill, C., Bell, S. M., Huddleston, J., Potter, B., Callender, C., Sagulenko, P., Bedford, T., & Neher, R. A. (2018). Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics*, *34*(23), 4121-4123.

Harder, C. B., Persson, S., Christensen, J., Ljubic, A., Nielsen, E. M., & Hoorfar, J. (2021). Molecular diagnostics of and in human/animal fecal samples remain feasible after long-term sample storage without specific

requirements. *Aims Microbiology*, *7*(4), 399-414. https://doi.org/10.3934/microbiol.2021024

Hilt, E. E., & Ferrieri, P. (2022). Next Generation and Other Sequencing Technologies in Diagnostic Microbiology and Infectious Diseases. *Genes*, *13*(9). https://doi.org/ARTN 1566
10.3390/genes13091566

Holt, K. E., Teo, Y. Y., Li, H., Nair, S., Dougan, G., Wain, J., & Parkhill, J. (2009). Detecting SNPs and estimating allele frequencies in clonal bacterial populations by sequencing pooled DNA. *Bioinformatics*, *25*(16), 2074-2075.

Horan, K., Da Silva, A. G., & Perry, A. (2022). MDU-PHL/abritamr: Update DB and publish. *Zenodo*. https://doi.org/https://doi.org/10.5281/ZENODO.7370627

Hormeño, L., Campos, M. J., Vadillo, S., & Quesada, A. (2020a). Occurrence of (O/M/O) Mosaic Gene in Tetracycline-Resistant. *Microorganisms*, *8*(11). https://doi.org/ARTN 1710
10.3390/microorganisms8111710

Hormeño, L., Campos, M. J., Vadillo, S., & Quesada, A. (2020b). Occurrence of tet(O/M/O) Mosaic Gene in Tetracycline-Resistant *Campylobacter*. *Microorganisms*, *8*(11). https://doi.org/ARTN 1710
10.3390/microorganisms8111710

Huang, A. D., Luo, C., Pena-Gonzalez, A., Weigand, M. R., Tarr, C. L., & Konstantinidis, K. T. (2017). Metagenomics of two severe foodborne outbreaks provides diagnostic signatures and signs of coinfection not attainable by traditional methods. *Applied and Environmental Microbiology*, *83*(3), e02577-02516.

Huang, Y., Leming, C. L., Suyemoto, M., & Altier, C. (2007). Genome-wide screen of *Salmonella* genes expressed during infection in pigs, using in vivo expression technology. *Appl Environ Microbiol*, *73*(23), 7522-7530. https://doi.org/10.1128/AEM.01481-07

Huisman, J. S. S., Vaughan, T. G. G., Egli, A., Tschudin-Sutter, S., Stadler, T., & Bonhoeffer, S. (2022). The effect of sequencing and assembly on the inference of horizontal gene transfer on chromosomal and plasmid phylogenies. *Philosophical Transactions of the Royal Society B-Biological Sciences*, *377*(1861). https://doi.org/ARTN 20210245
10.1098/rstb.2021.0245

Huptas, C., Scherer, S., & Wenning, M. (2016). Optimized Illumina PCR-free library preparation for bacterial whole genome sequencing and analysis of factors influencing de novo assembly. *BMC research notes*, *9*, 1-14.

Hyatt, D., Chen, G. L., Locascio, P. F., Land, M. L., Larimer, F. W., & Hauser, L. J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, *11*, 119. https://doi.org/10.1186/1471-2105-11-119

Inns, T., Ashton, P. M., Herrera-Leon, S., Lighthill, J., Foulkes, S., Jombart, T., Rehman, Y., Fox, A., Dallman, T., De Pinna, E., Browning, L., Coia, J. E., Edeghere, O., & Vivancos, R. (2017). Prospective use of whole genome sequencing (WGS) detected a multi-country outbreak of *Salmonella*

Enteritidis. *Epidemiology and Infection*, *145*(2), 289-298. https://doi.org/10.1017/S0950268816001941

Isokääntä, H., Tomnikov, N., Vanhatalo, S., Munukka, E., Huovinen, P., Hakanen, A. J., & Kallonen, T. (2024). High-throughput DNA extraction strategy for fecal microbiome studies. *Microbiology Spectrum*, *12*(6). https://doi.org/10.1128/spectrum.02932-23

Jacob, J. J., Pragasam, A. K., Vasudevan, K., Veeraraghavan, B., Kang, G., John, J., Nagvekar, V., & Mutreja, A. (2021). Typhi acquires diverse plasmids from other *Enterobacteriaceae* to develop cephalosporin resistance. *Genomics*, *113*(4), 2171-2176. https://doi.org/10.1016/j.ygeno.2021.05.003

Jaudou, S., Tran, M. L., Vorimore, F., Fach, P., & Delannoy, S. (2022). Evaluation of high molecular weight DNA extraction methods for long-read sequencing of Shiga toxin-producing *Escherichia coli*. *Plos One*, *17*(7). https://doi.org/ARTN e0270751
10.1371/journal.pone.0270751

Joensen, K. G., Kiil, K., Gantzhorn, M. R., Nauerby, B., Engberg, J., Holt, H. M., Nielsen, H. L., Petersen, A. M., Kuhn, K. G., Sando, G., Ethelberg, S., & Nielsen, E. M. (2020). Whole-Genome Sequencing to Detect Numerous *Campylobacter jejuni* Outbreaks and Match Patient Isolates to Sources, Denmark, 2015–2017. *Emerging Infectious Diseases*, *26*(3), 523-532. https://doi.org/10.3201/eid2603.190947

Joensen, K. G., Schjorring, S., Gantzhorn, M. R., Vester, C. T., Nielsen, H. L., Engberg, J. H., Holt, H. M., Ethelberg, S., Müller, L., Sando, G., & Nielsen, E. M. (2021). Whole genome sequencing data used for surveillance of *Campylobacter* infections: detection of a large continuous outbreak, Denmark, 2019. *Eurosurveillance*, *26*(22). https://doi.org/Artn 200139
10.2807/1560-7917.Es.2021.26.22.2001396

Joseph, L. A., Griswold, T., Vidyaprakash, E., Im, S. B., Williams, G. M., Pouseele, H. A., Hise, K. B., & Carleton, H. A. (2023). Evaluation of core genome and whole genome multilocus sequence typing schemes for *Campylobacter jejuni* and *Campylobacter coli* outbreak detection in the USA. *Microbial Genomics*, *9*(5). https://doi.org/ARTN 001012
10.1099/mgen.0.001012

Juraschek, K., Borowiak, M., Tausch, S. H., Malorny, B., Käsbohrer, A., Otani, S., Schwarz, S., Meemken, D., Deneke, C., & Hammerl, J. A. (2021). Outcome of different sequencing and assembly approaches on the detection of plasmids and localization of antimicrobial resistance genes in commensal *Escherichia coli*. *Microorganisms*, *9*(3), 598.

Kaakoush, N. O., Castaño-Rodríguez, N., Mitchell, H. M., & Man, S. I. M. (2015). Global epidemiology of *Campylobacter* infection. *Clinical Microbiology Reviews*, *28*(3), 687-720. https://doi.org/10.1128/Cmr.00006-15

Kazantseva, J., Malv, E., Kaleda, A., Kallastu, A., & Meikas, A. (2021). Optimisation of sample storage and DNA extraction for human gut microbiota studies. *Bmc Microbiology*, *21*(1). https://doi.org/ARTN 158
10.1186/s12866-021-02233-y

Kenyon, J., Inns, T., Aird, H., Swift, C., Astbury, J., Forester, E., & Decraene. (2020). *Campylobacter* outbreak associated with raw drinking milk, North West England, 2016. *Epidemiology & Infection*, *148*. https://doi.org/ARTN e13
10.1017/S0950268820000096
Khattak, F., Galgano, S., & Houdijk, J. (2022). Bacterial concentration and *Campylobacter* spp. quantification differ when fresh or ultra-frozen samples are analysed over time using molecular biology and culture-based methods. *Plos One*, *17*(9). https://doi.org/ARTN e0274682
10.1371/journal.pone.0274682
Khezri, A., Avershina, E., & Ahmad, R. (2021). Hybrid assembly provides improved resolution of plasmids, antimicrobial resistance genes, and virulence factors in *Escherichia coli* and *Klebsiella pneumoniae* clinical isolates. *Microorganisms*, *9*(12), 2560.
Kim, C., Pongpanich, M., & Porntaveetus, T. (2024). Unraveling metagenomics through long-read sequencing: a comprehensive review. *Journal of Translational Medicine*, *22*(1), 111.
Kim, D., Song, L., Breitwieser, F. P., & Salzberg, S. L. (2016). Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome Research*, *26*(12), 1721-1729. https://doi.org/10.1101/gr.210641.116
Kim, J. I., Maguire, F., Tsang, K. K., Gouliouris, T., Peacock, S. J., McAllister, T. A., McArthur, A. G., & Beiko, R. G. (2022). Machine Learning for Antimicrobial Resistance Prediction: Current Practice, Limitations, and Clinical Perspective. *Clinical Microbiology Reviews*, *35*(3). https://doi.org/10.1128/cmr.00179-21
Kim, J. S., Liu, L., Kant, S., Orlicky, D. J., Uppalapati, S., Margolis, A., Davenport, B. J., Morrison, T. E., Matsuda, J., Mcclelland, M., Jones-Carson, J., & Vazquez-Torres, A. (2024). Anaerobic respiration of host-derived methionine sulfoxide protects intracellular *Salmonella* from the phagocyte NADPH oxidase. *Cell Host & Microbe*, *32*(3). https://doi.org/10.1016/j.chom.2024.01.004
Ko, K. K. K., Chng, K. R., & Nagarajan, N. (2022). Metagenomics-enabled microbial surveillance. *Nature Microbiology*, *7*(4), 486-496. https://doi.org/10.1038/s41564-022-01089-w
Koren, S., & Phillippy, A. M. (2015). One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly. *Current opinion in microbiology*, *23*, 110-120.
Köser, C. U., Ellington, M. J., Cartwright, E. J., Gillespie, S. H., Brown, N. M., Farrington, M., Holden, M. T., Dougan, G., Bentley, S. D., & Parkhill, J. (2012). Routine use of microbial whole genome sequencing in diagnostic and public health microbiology.
Kozyreva, V. K., Crandall, J., Sabol, A., Poe, A., Zhang, P., Concepción-Acevedo, J., Schroeder, M. N., Wagner, D., Higa, J., & Trees, E. (2016). Laboratory investigation of *Salmonella enterica* serovar Poona outbreak in California: comparison of pulsed-field gel electrophoresis (PFGE) and whole genome sequencing (WGS) results. *PLoS currents*, *8*, ecurrents. outbreaks. 1bb3e36e74bd5779bc5743ac5773a5778dae5752e5776.

Krawczyk, P. S., Lipinski, L., & Dziembowski, A. (2018). PlasFlow: predicting plasmid sequences in metagenomic data using genome signatures. *Nucleic Acids Research*, *46*(6). https://doi.org/ARTN e35
10.1093/nar/gkx1321

Kruasuwan, W., Sawatwong, P., Jenjaroenpun, P., Wankaew, N., Arigul, T., Yongkiettrakul, S., Lunha, K., Sudjai, A., Siludjai, D., & Skaggs, B. (2024). Comparative evaluation of commercial DNA isolation approaches for nanopore-only bacterial genome assembly and plasmid recovery. *Scientific reports*, *14*(1), 27672.

Kubota, K. A., Wolfgang, W. J., Baker, D. J., Boxrud, D., Turner, L., Trees, E., Carleton, H. A., & Gerner-Smidt, P. (2019). PulseNet and the Changing Paradigm of Laboratory-Based Surveillance for Foodborne Diseases. *Public Health Reports*, *134*(2_Suppl), 22s-28s. https://doi.org/10.1177/0033354919881650

Kumburu, H. H., Shayo, M., van Zwetslaar, M., Njau, J., Kuchaka, D. J., Ignas, I. P., Wadugu, B., Kasworm, R., Masaki, L. J., & Hallgren, M. B. (2023). Nanopore sequencing technology for clinical diagnosis of infectious diseases where laboratory capacity is meager: A case report. *Heliyon*, *9*(7).

Kwa, W. T., Sim, C. K., Low, A., & Lee, J. W. J. (2024). A Comparison of Three Automated Nucleic Acid Extraction Systems for Human Stool Samples. *Microorganisms*, *12*(12). https://doi.org/ARTN 2417
10.3390/microorganisms12122417

Kwon, H. J., Chen, Z., Evans, P., Meng, J. H., & Chen, Y. (2020). Characterization of Mobile Genetic Elements Using Long-Read Sequencing for Tracking *Listeria monocytogenes* from Food Processing Environments. *Pathogens*, *9*(10). https://doi.org/ARTN 822
10.3390/pathogens9100822

Kwong, J. C., McCallum, N., Sintchenko, V., & Howden, B. P. (2015). Whole genome sequencing in clinical and public health microbiology. *Pathology*, *47*(3), 199-210.

Laisnez, V., Vusirikala, A., Nielsen, C. S., Cantaert, V., Delbrassinne, L., Mattheus, W., Verhaegen, B., Delamare, H., Jourdan-Da Silva, N., Lachmann, R., Simon, S., Cormican, M., Garvey, P., McKeown, P., Stephan, R., Brown, D., Browning, L., Hoban, A., Larkin, L., . . . Van Cauteren, D. (2025). Key role of whole genome sequencing in resolving an international outbreak of monophasic Salmonella Typhimurium linked to chocolate products. *BMC Infect Dis*, *25*(1), 242. https://doi.org/10.1186/s12879-025-10629-8

Lamas, A., Miranda, J. M., Regal, P., Vázquez, B., Franco, C. M., & Cepeda, A. (2018). A comprehensive review of non-enterica subspecies of *Salmonella enterica*. *Microbiological Research*, *206*, 60-73. https://doi.org/10.1016/j.micres.2017.09.010

Landman, F., Jamin, C., de Haan, A., Witteveen, S., Bos, J., van der Heide, H. G., Schouls, L. M., & Hendrickx, A. P. (2024). Genomic surveillance of multidrug-resistant organisms based on long-read sequencing. *Genome Medicine*, *16*(1), 137.

Langridge, G. C., Wain, J., & Nair, S. (2012). Invasive Salmonellosis in Humans. *EcoSal Plus*, *5*(1). https://doi.org/10.1128/ecosalplus.8.6.2.2

Lapidus, A. L., & Korobeynikov, A. I. (2021). Metagenomic data assembly–the way of decoding unknown microorganisms. *Frontiers in Microbiology*, *12*, 613791.

Leblanc-Maridor, M., Beaudeau, F., Seegers, H., Denis, M., & Belloc, C. (2011). Rapid identification and quantification of *Campylobacter coli* and *Campylobacter jejuni* by real-time PCR in pure cultures and in complex samples. *Bmc Microbiology*, *11*, 1-16.

LeFrançois, B., & Cunningham, L. (2019). Evaluation of DNA extraction methods to obtain accurate and reliable results from gut microbiome samples. *DNA Genotek*.

Lever, M. A., Torti, A., Eickenbusch, P., Michaud, A. B., Šantl-Temkiv, T., & Jørgensen, B. B. (2015). A modular method for the extraction of DNA and RNA, and the separation of DNA pools from diverse environmental sample types. *Frontiers in Microbiology*, *6*, 476.

Li, C., Tyson, G. H., Hsu, C. H., Harrison, L., Strain, E., Tran, T. T., Tillman, G. E., Dessai, U., McDermott, P. F., & Zhao, S. H. (2021). Long-Read Sequencing Reveals Evolution and Acquisition of Antimicrobial Resistance and Virulence Genes in *Salmonella enterica*. *Frontiers in Microbiology*, *12*. https://doi.org/ARTN 777817
10.3389/fmicb.2021.777817

Li, D., He, S., Dong, R., Cui, Y., & Shi, X. (2022). Stress Response Mechanisms of *Salmonella* Enteritidis to Sodium Hypochlorite at the Proteomic Level. *Foods*, *11*(18). https://doi.org/10.3390/foods11182912

Li, D., Liu, C. M., Luo, R., Sadakane, K., & Lam, T. W. (2015). MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics*, *31*(10), 1674-1676. https://doi.org/10.1093/bioinformatics/btv033

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, *34*(18), 3094-3100. https://doi.org/10.1093/bioinformatics/bty191

Li, W., Cui, Q., Bai, L., Fu, P., Han, H., Liu, J., & Guo, Y. (2021). Application of whole-genome sequencing in the national molecular tracing network for foodborne disease surveillance in China. *Foodborne pathogens and disease*, *18*(8), 538-546.

Li, X. M., Shi, X., Yao, Y., Shen, Y. C., Wu, X. L., Cai, T., Liang, L. X., & Wang, F. (2023). Effects of Stool Sample Preservation Methods on Gut Microbiota Biodiversity: New Original Data and Systematic Review with Meta-Analysis. *Microbiology Spectrum*, *11*(3). https://doi.org/10.1128/spectrum.04297-22

Libuit, K. G., Doughty, E. L., Otieno, J. R., Ambrosio, F., Kapsak, C. J., Smith, E. A., Wright, S. M., Scribner, M. R., Petit, R. I. I. I., Mendes, C. I., Huergo, M., Legacki, G., Loreth, C., Park, D. J., & Sevinsky, J. R. (2023). Accelerating bioinformatics implementation in public health. *Microbial Genomics*, *9*(7). https://doi.org/ARTN 001051
10.1099/mgen.0.001051

Lieberman, T. D., Flett, K. B., Yelin, I., Martin, T. R., McAdam, A. J., Priebe, G. P., & Kishony, R. (2014). Genetic variation of a bacterial pathogen within individuals with cystic fibrosis provides a record of selective pressures. *Nature genetics*, *46*(1), 82-87.

Lin, Y., Yuan, J., Kolmogorov, M., Shen, M. W., Chaisson, M., & Pevzner, P. A. (2016). Assembly of long error-prone reads using de Bruijn graphs. *Proceedings of the national academy of sciences*, *113*(52), E8396-E8405.

Lipkin, W. I. (2010). Microbe Hunting. *Microbiology and Molecular Biology Reviews*, *74*(3), 363-+. https://doi.org/10.1128/Mmbr.00007-10

Liu, K., Burke, J., Kilburn, D., Fedak, R., Kim, M., Ferrao, H., Galvin, B., Bjornson, K., Workman, R., & Hilger, N. (2019). High-Throughput, High MW DNA Extraction and Size Selection for Long-Read Sequencing. *Journal of Biomolecular Techniques: JBT*, *30*(Suppl), S19.

Liu, Q., Via, L., Luo, T., Liang, L., Liu, X., Wu, S., Shen, Q., Wei, W., Ruan, X., & Yuan, X. (2015). Within patient microevolution of *Mycobacterium tuberculosis* correlates with heterogeneous responses to treatment. *Scientific reports*, *5*(1), 17507.

Loman, N. J., Constantinidou, C., Christner, M., Rohde, H., Chan, J. Z. M., Quick, J., Weir, J. C., Quince, C., Smith, G. P., Betley, J. R., Aepfelbacher, M., & Pallen, M. J. (2013). A culture-independent sequence-based metagenomics approach to the investigation of an outbreak of Shiga-toxigenic *Escherichia coli* O104: H4. *Jama-Journal of the American Medical Association*, *309*(14), 1502-1510. https://doi.org/10.1001/jama.2013.3231

Lou, R. N., Jacobs, A., Wilder, A. P., & Therkildsen, N. O. (2021). A beginner's guide to low-coverage whole genome sequencing for population genomics. *Molecular ecology*, *30*(23), 5966-5993.

Lu, J., Rincon, N., Wood, D. E., Breitwieser, F. P., Pockrandt, C., Langmead, B., Salzberg, S. L., & Steinegger, M. (2022). Metagenome analysis using the Kraken software suite. *Nature Protocols*, *17*(12), 2815-2839. https://doi.org/10.1038/s41596-022-00738-y

Luan, T., Commichaux, S., Hoffmann, M., Jayeola, V., Jang, J. H., Pop, M., Rand, H., & Luo, Y. (2024). Benchmarking short and long read polishing tools for nanopore assemblies: achieving near-perfect genomes for outbreak isolates. *Bmc Genomics*, *25*(1). https://doi.org/ARTN 679
10.1186/s12864-024-10582-x

Maghini, D. G., Moss, E. L., Vance, S. E., & Bhatt, A. S. (2021). Improved high-molecular-weight DNA extraction, nanopore sequencing and metagenomic assembly from the human gut microbiome. *Nature Protocols*, *16*(1), 458-471. https://doi.org/10.1038/s41596-020-00424-x

Mai, Q. D., Luo, Y. S., Ye, R. T., Jiang, Y. Y., Qin, Y. T., Guo, J. F., Lai, W. M., Wu, Y. B., & Luo, M. Y. (2025). Detection of *Salmonella* spp. Related Co-Infections Among Children with Diarrheal Diseases in Guangzhou, China. *Infection and Drug Resistance*, *18*, 1895-1903. https://doi.org/10.2147/Idr.S515033

Mäklin, T., Kallonen, T., Alanko, J., Samuelsen, Ø., Hegstad, K., Mäkinen, V., Corander, J., Heinz, E., & Honkela, A. (2021). Bacterial genomic epidemiology with mixed samples. *Microbial Genomics*, *7*(11), 000691.

Man, S. M. (2011). The clinical importance of emerging *Campylobacter* species. *Nature Reviews Gastroenterology & Hepatology*, *8*(12), 669-685. https://doi.org/10.1038/nrgastro.2011.191

Mank, L., Mandour, M., Rabatsky-Ehr, T., Phan, Q., Krasnitski, J., Brockmeyer, J., Bushnell, L., Applewhite, C., Cartter, M., & Kattan, J. (2010). Multiple-Serotype *Salmonella* Gastroenteritis Outbreak After a Reception--Connecticut, 2009. *MMWR: Morbidity & Mortality Weekly Report*, *59*(34).

Marchello, C. S., Birkhold, M., & Crump, J. A. (2020). Complications and mortality of typhoid fever: a global systematic review and meta-analysis. *Journal of Infection*, *81*(6), 902-910.

Martins, L. D., Page, A. J., Mather, A. E., & Charles, I. G. (2020). Taxonomic resolution of the ribosomal RNA operon in bacteria: implications for its use with long-read sequencing. *Nar Genomics and Bioinformatics*, *2*(1). https://doi.org/ARTN lqz016
10.1093/nargab/lqz016

McClelland, M., Sanderson, K. E., Clifton, S. W., Latreille, P., Porwollik, S., Sabo, A., Meyer, R., Bieri, T., Ozersky, P., McLellan, M., Harkins, C. R., Wang, C. Y., Nguyen, C., Berghoff, A., Elliott, G., Kohlberg, S., Strong, C., Du, F. Y., Carter, J., . . . Wilson, R. K. (2004). Comparison of genome degradation in Paratyphi A and Typhi, human-restricted serovars of *Salmonella enterica* that cause typhoid. *Nature genetics*, *36*(12), 1268-1274. https://doi.org/10.1038/ng1470

McGaughey, K. D., Yilmaz-Swenson, T., Elsayed, N. M., Cruz, D. A., Rodriguez, R. M., Kritzer, M. D., Peterchev, A. V., Gray, M., Lewis, S. R., Roach, J., Wetsel, W. C., & Williamson, D. E. (2019). Comparative evaluation of a new magnetic bead-based DNA extraction method from fecal samples for downstream next-generation 16S rRNA gene sequencing. *Plos One*, *14*(2). https://doi.org/ARTN e0212712
10.1371/journal.pone.0212712

Mehra, P., & Kumar, A. (2024). Emerging importance of stool preservation methods in OMICS studies with special focus on cancer biology. *Cell Biochemistry and Function*, *42*(5). https://doi.org/ARTN e4063
10.1002/cbf.4063

Meumann, E. M., Krause, V. L., Baird, R., & Currie, B. J. (2022). Using genomics to understand the epidemiology of infectious diseases in the Northern territory of Australia. *Tropical Medicine and Infectious Disease*, *7*(8), 181.

Meysman, P., Sanchez-Rodriguez, A., Fu, Q., Marchal, K., & Engelen, K. (2013). Expression divergence between *Escherichia coli* and *Salmonella enterica* serovar Typhimurium reflects their lifestyles. *Molecular biology and evolution*, *30*(6), 1302-1314.

Miller, K. A., Phillips, R. S., Mrázek, J., & Hoover, T. R. (2013). *Salmonella* utilizes D-glucosaminate via a mannose family phosphotransferase system

permease and associated enzymes. *Journal of bacteriology*, *195*(18), 4057-4066.

Mills, C. K., & Gherna, R. L. (1988). Cryopreservation Studies of *Campylobacter*. *Cryobiology*, *25*(2), 148-152. https://doi.org/Doi 10.1016/0011-2240(88)90008-9

Mixao, V., Pinto, M., Brendebach, H., Sobral, D., Santos, J. D., Radomski, N., Uldall, A. S. M., Bomba, A., Pietsch, M., Bucciacchio, A., de Ruvo, A., Castelli, P., Iwan, E., Simon, S., Coipan, C. E., Linde, J., Petrovska, L., Kaas, R. S., Joensen, K. G., . . . Borges, V. (2025). Multi-country and intersectoral assessment of cluster congruence between pipelines for genomics surveillance of foodborne pathogens. *Nature Communications*, *16*(1). https://doi.org/ARTN 3961 10.1038/s41467-025-59246-8

Mook, P., Gardiner, D., Verlander, N. Q., McCormick, J., Usdin, M., Crook, P., Jenkins, C., & Dallman, T. J. (2018). Operational burden of implementing *Salmonella* Enteritidis and Typhimurium cluster detection using whole genome sequencing surveillance data in England: a retrospective assessment. *Epidemiology and Infection*, *146*(11), 1452-1460. https://doi.org/10.1017/S0950268818001589

Moore, N. E., Wang, J., Hewitt, J., Croucher, D., Williamson, D. A., Paine, S., Yen, S. H., Greening, G. E., & Hall, R. J. (2015). Metagenomic Analysis of Viruses in Feces from Unsolved Outbreaks of Gastroenteritis in Humans. *Journal of Clinical Microbiology*, *53*(1), 15-21. https://doi.org/10.1128/Jcm.02029-14

Moss, E. L., Maghini, D. G., & Bhatt, A. S. (2020). Complete, closed bacterial genomes from microbiomes using nanopore sequencing. *Nature Biotechnology*, *38*(6), 701-+. https://doi.org/10.1038/s41587-020-0422-6

Mu, S. R., Hu, L., Zhang, Y., Liu, Y. M., Cui, X. J., Zou, X. H., Wang, Y. M., Lu, B. H., Zhou, S. L., Liang, X. X., Liang, C., Xiao, N., O'Grady, J., Lee, S., & Cao, B. (2021). Prospective Evaluation of a Rapid Clinical Metagenomics Test for Bacterial Pneumonia. *Frontiers in Cellular and Infection Microbiology*, *11*. https://doi.org/ARTN 684965 10.3389/fcimb.2021.684965

Nadkarni, M. A., Martin, F. E., Hunter, N., & Jacques, N. A. (2009). Methods for optimizing DNA extraction before quantifying oral bacterial numbers by real-time PCR. *FEMS microbiology letters*, *296*(1), 45-51.

Nair, S., Chattaway, M., Langridge, G. C., Gentle, A., Day, M., Ainsworth, E. V., Mohamed, I., Smith, R., Jenkins, C., Dallman, T. J., & Godbole, G. (2021). ESBL-producing strains isolated from imported cases of enteric fever in England and Wales reveal multiple chromosomal integrations of blaCTX-M-15 in XDR *Salmonella* Typhi. *Journal of antimicrobial chemotherapy*, *76*(6), 1459-1466. https://doi.org/10.1093/jac/dkab049

Natarajan, A., Han, A., Zlitni, S., Brooks, E. F., Vance, S. E., Wolfe, M., Singh, U., Jagannathan, P., Pinsky, B. A., Boehm, A., & Bhatt, A. S. (2021). Standardized preservation, extraction and quantification techniques for detection of fecal SARS-CoV-2 RNA (vol 12, 5753, 2021). *Nature Communications*, *12*(1). https://doi.org/ARTN 7100

10.1038/s41467-021-27392-4
Neal-McKinney, J. M., Liu, K. C., Lock, C. M., Wu, W. H., & Hu, J. (2021). Comparison of MiSeq, MinION, and hybrid genome sequencing for analysis of *Campylobacter jejuni*. *Scientific reports*, *11*(1). https://doi.org/ARTN 5676
10.1038/s41598-021-84956-6
Nel Van Zyl, K., Whitelaw, A. C., & Newton-Foot, M. (2020). The effect of storage conditions on microbial communities in stool. *Plos One*, *15*(1), e0227486.
Neuert, S., Nair, S., Day, M. R., Doumith, M., Ashton, P. M., Mellor, K. C., Jenkins, C., Hopkins, K. L., Woodford, N., de Pinna, E., Godbole, G., & Dallman, T. J. (2018). Prediction of Phenotypic Antimicrobial Resistance Profiles From Whole Genome Sequences of Non-typhoidal *Salmonella enterica*. *Frontiers in Microbiology*, *9*. https://doi.org/ARTN 592
10.3389/fmicb.2018.00592
Newland, G. A. I., Gibson, G. R., Jackson, F. L., & Wijeyesekera, A. (2021). Assessment of stool collection and storage conditions for in vitro human gut model studies. *Journal of Microbiological Methods*, *185*. https://doi.org/ARTN 106230
10.1016/j.mimet.2021.106230
Nicholls, S. M., Poplawski, R., Bull, M. J., Underwood, A., Chapman, M., Abu-Dahab, K., Taylor, B., Colquhoun, R. M., Rowe, W. P. M., Jackson, B., Hill, V., O'Toole, A., Rey, S., Southgate, J., Amato, R., Livett, R., Gonçalves, S., Harrison, E. M., Peacock, S. J., . . . Cons, C.-G. U. C.-U. (2021). CLIMB-COVID: continuous integration supporting decentralised sequencing for SARS-CoV-2 genomic surveillance. *Genome Biology*, *22*(1). https://doi.org/ARTN 196
10.1186/s13059-021-02395-y
Noyes, N. R., Yang, X., Linke, L. M., Magnuson, R. J., Cook, S. R., Zaheer, R., Yang, H., Woerner, D. R., Geornaras, I., & McArt, J. A. (2016). Characterization of the resistome in manure, soil and wastewater from dairy and beef production systems. *Scientific reports*, *6*(1), 24645.
Nursofiah, S., Hartoyo, Y., Amalia, N., Febrianti, T., Febriyana, D., Saraswati, R., Puspandari, N., Sariadji, K., Rukminiati, Y., & Muna, F. (2021). Long-term storage of bacterial isolates by using tryptic Soy Broth with 15% glycerol in the deep freezer (-70 to-80 C). IOP Conference Series: Earth and Environmental Science,
O'Sullivan, V., Madrid-Gambin, F., Alegra, T., Gibbons, H., & Brennan, L. (2018). Impact of Sample Storage on the NMR Fecal Water Metabolome. *Acs Omega*, *3*(12), 16585-16590. https://doi.org/10.1021/acsomega.8b01761
Olm, M. R., Crits-Christoph, A., Bouma-Gregson, K., Firek, B. A., Morowitz, M. J., & Banfield, J. F. (2021). inStrain profiles population microdiversity from metagenomic data and sensitively detects shared microbial strains. *Nature Biotechnology*, *39*(6), 727-736.
Olson, N. D., Treangen, T. J., Hill, C. M., Cepeda-Espinoza, V., Ghurye, J., Koren, S., & Pop, M. (2019). Metagenomic assembly through the lens of

validation: recent advances in assessing and improving the quality of genomes assembled from metagenomes. *Briefings in bioinformatics*, *20*(4), 1140-1150.

Oswald, J., Njenga, R., Natriashvili, A., Sarmah, P., & Koch, H.-G. (2021). The dynamic SecYEG translocon. *Frontiers in molecular biosciences*, *8*, 664241.

Pacific Biosciences. (2022). *Preparing DNA for PacBio HiFi Sequencing — Extraction and Quality Control (Technical Note)*. https://www.pacb.com/wp-content/uploads/Technical-Note-Preparing-DNA-for-PacBio-HiFi-Sequencing-Extraction-and-Quality-Control.pdf

Page, A. J., Ainsworth, E. V., & Langridge, G. C. (2020). *socru*: typing of genome-level order and orientation around ribosomal operons in bacteria. *Microbial Genomics*, *6*(7). https://doi.org/ARTN 000396
10.1099/mgen.0.000396

Page, A. J., Alikhan, N. F., Carleton, H. A., Seemann, T., Keane, J. A., & Katz, L. S. (2017). Comparison of classical multi-locus sequence typing software for next-generation sequencing data. *Microbial Genomics*, *3*(8). https://doi.org/ARTN 000124
10.1099/mgen.0.000124

Pan, S., Zhu, C., Zhao, X. M., & Coelho, L. P. (2022). A deep siamese neural network improves metagenome-assembled genomes in microbiome datasets across different environments. *Nature Communications*, *13*(1), 2326. https://doi.org/10.1038/s41467-022-29843-y

Panek, M., Paljetak, H. C., Baresic, A., Peric, M., Matijasic, M., Lojkic, I., Bender, D. V., Krznaric, Z., & Verbanac, D. (2018). Methodology challenges in studying human gut microbiota - effects of collection, storage, DNA extraction and next generation sequencing technologies. *Scientific reports*, *8*. https://doi.org/ARTN 5143
10.1038/s41598-018-23296-4

Park, C. J., Li, J. F., Zhang, X. L., Gao, F. X., Benton, C. S., & Andam, C. P. (2020). Genomic Epidemiology and Evolution of Diverse Lineages of Clinical Cocirculating in New Hampshire, USA, 2017. *Journal of Clinical Microbiology*, *58*(6). https://doi.org/ARTN e02070-19
10.1128/JCM.02070-19

Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P., & Tyson, G. W. (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research*, *25*(7), 1043-1055. https://doi.org/10.1101/gr.186072.114

Paulos, S., Mateo, M., de Lucio, A., Hernández-de Mingo, M., Bailo, B., Saugar, J. M., Cardona, G. A., Fuentes, I., Mateo, M., & Carmena, D. (2016). Evaluation of five commercial methods for the extraction and purification of DNA from human faecal samples for downstream molecular detection of the enteric protozoan parasites *Cryptosporidium* spp., *Giardia duodenalis*, and *Entamoeba* spp. *Journal of Microbiological Methods*, *127*, 68-73. https://doi.org/10.1016/j.mimet.2016.05.020

Pearce, M. E., Langridge, G. C., Lauer, A., Grant, K., Maiden, M. C., & Chattaway, M. A. (2021). An evaluation of the species and subspecies of

the genus Salmonella with whole genome sequence data: Proposal of type strains and epithets for novel S. enterica subspecies VII, VIII, IX, X and XI. *Genomics*, *113*(5), 3152-3162.

Peng, J., Feng, J., Ji, H., Kong, X., Hong, J., Zhu, L., & Qian, H. (2024). Emergence of Rarely Reported Extensively Drug-Resistant Salmonella Enterica Serovar Paratyphi B among Patients in East China. *Antibiotics*, *13*(6), 519.

Pereira-Marques, J., Hout, A., Ferreira, R. M., Weber, M., Pinto-Ribeiro, I., van Doorn, L. J., Knetsch, C. W., & Figueiredo, C. (2019). Impact of Host DNA and Sequencing Depth on the Taxonomic Resolution of Whole Metagenome Sequencing for Microbiome Analysis. *Frontiers in Microbiology*, *10*. https://doi.org/ARTN 1277
10.3389/fmicb.2019.01277

Pérez-Losada, M., Cabezas, P., Castro-Nallar, E., & Crandall, K. A. (2013). Pathogen typing in the genomics era: MLST and the future of molecular epidemiology. *Infection, Genetics and Evolution*, *16*, 38-53.

Perrocheau, A., Jephcott, F., Asgari-Jirhanden, N., Greig, J., Peyraud, N., & Tempowski, J. (2023). Investigating outbreaks of initially unknown aetiology in complex settings: findings and recommendations from 10 case studies. *International Health*, *15*(5), 537-546. https://doi.org/10.1093/inthealth/ihac088

Peterson, C. L., Alexander, D., Chen, J. C. Y., Adam, H., Walker, M., Ali, J., Forbes, J., Taboada, E., Barker, D. O. R., Graham, M., Knox, N., & Reimer, A. R. (2022). Clinical Metagenomics Is Increasingly Accurate and Affordable to Detect Enteric Bacterial Pathogens in Stool. *Microorganisms*, *10*(2). https://doi.org/ARTN 441
10.3390/microorganisms10020441

Polonis, K., Blommel, J. H., Hughes, A. E., Spencer, D., Thompson, J. A., & Schroeder, M. C. (2025). Innovations in Short-Read sequencing technologies and their applications to clinical genomics. *Clinical chemistry*, *71*(1), 97-108.

Popa, G. L., & Popa, M. I. (2021). *Salmonella* spp. infection - a continuous threat worldwide. *Germs*, *11*(1), 88-96. https://doi.org/10.18683/germs.2021.1244

Promega. (n.d.). *DNA purification: DNA extraction methods*. Promega Corporation. Retrieved 2 March from https://www.promega.co.uk/resources/guides/nucleic-acid-analysis/dna-purification

Purushothaman, S., Meola, M., Roloff, T., Rooney, A. M., & Egli, A. (2024). Evaluation of DNA extraction kits for long-read shotgun metagenomics using Oxford Nanopore sequencing for rapid taxonomic and antimicrobial resistance detection. *Scientific reports*, *14*(1). https://doi.org/ARTN 29531
10.1038/s41598-024-80660-3

Quick, J., & Loman, N. J. (2019). DNA extraction strategies for nanopore sequencing. *Nanopore sequencing: An introduction*, 1-17.

Raghuram, V., Gunoskey, J. J., Hofstetter, K. S., Jacko, N. F., Shumaker, M. J., Hu, Y.-J., Read, T. D., & David, M. Z. (2023). Comparison of genomic diversity between single and pooled Staphylococcus aureus colonies isolated from human colonization cultures. *Microbial Genomics*, *9*(11), 001111.

Reuter, T., & Zaheer, R. (2016). Nucleic Acid Sample Preparation from Feces and Manure. *Sample Preparation Techniques for Soil, Plant, and Animal Samples*, 341-352.

Revez, J., Espinosa, L., Albiger, B., Leitmeyer, K. C., Struelens, M. J., Points, E. N. M. F., & Group, E. (2017). Survey on the use of whole-genome sequencing for infectious diseases surveillance: rapid expansion of European national capacities, 2015–2016. *Frontiers in public health*, *5*, 347.

Ribarska, T., Bjornstad, P. M., Sundaram, A. Y. M., & Gilfillan, G. D. (2022). Optimization of enzymatic fragmentation is crucial to maximize genome coverage: a comparison of library preparation methods for Illumina sequencing. *Bmc Genomics*, *23*(1). https://doi.org/ARTN 92 10.1186/s12864-022-08316-y

Ribot, E. M., Freeman, M., Hise, K. B., & Gerner-Smidt, P. (2019). PulseNet: entering the age of next-generation sequencing. *Foodborne pathogens and disease*, *16*(7), 451-456.

Rintarhat, P., Cho, Y.-J., Koh, H., Park, S., Lee, E. J., Lim, H., Noh, J., Lee, D.-W., & Jung, W. H. (2024). Assessment of DNA extraction methods for human gut mycobiome analysis. *Royal Society Open Science*, *11*(1), 231129.

Roopnarain, A., Mukhuba, M., Adeleke, R., & Moeletsi, M. (2017). Biases during DNA extraction affect bacterial and archaeal community profile of anaerobic digestion samples. *3 Biotech*, *7*(6), 375.

Rudder, S. J., Djeghout, B., Elumogo, N., Janecko, N., & Langridge, G. C. (2025). Genomic diversity of non-typhoidal Salmonella found in patients suffering from gastroenteritis in Norfolk, UK. *Microbial Genomics*, *11*(8), 001468.

Salter, S. J., Cox, M. J., Turek, E. M., Calus, S. T., Cookson, W. O., Moffatt, M. F., Turner, P., Parkhill, J., Loman, N. J., & Walker, A. W. (2014). Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *Bmc Biology*, *12*. https://doi.org/ARTN 87 10.1186/s12915-014-0087-z

Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA Sequencing with Chain-Terminating Inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, *74*(12), 5463-5467. https://doi.org/DOI 10.1073/pnas.74.12.5463

Santos, L. S., Rossi, D. A., Braz, R. F., Fonseca, B. B., Guidotti-Takeuchi, M., Alves, R. N., Beletti, M. E., Almeida-Souza, H. O., Maia, L. P., Santos, P. D., de Souza, J. B., & de Melo, R. T. (2023). Roles of viable but non-culturable state in the survival of *Campylobacter jejuni*. *Frontiers in Cellular and Infection Microbiology*, *13*. https://doi.org/ARTN 1122450 10.3389/fcimb.2023.1122450

Scarano, C., Veneruso, I., De Simone, R. R., Di Bonito, G., Secondino, A., & D'Argenio, V. (2024). The Third-Generation Sequencing Challenge: Novel Insights for the Omic Sciences. *Biomolecules*, *14*(5). https://doi.org/ARTN 568
10.3390/biom14050568
Schmid, M., Frei, D., Patrignani, A., Schlapbach, R., Frey, J. E., Remus-Emsermann, M. N. P., & Ahrens, C. H. (2018). Pushing the limits of de novo genome assembly for complex prokaryotic genomes harboring very long, near identical repeats *Nucleic Acids Research*, *46*(17), 8953-8965. https://doi.org/10.1093/nar/gky726
Sedlazeck, F. J., Rescheneder, P., Smolka, M., Fang, H., Nattestad, M., von Haeseler, A., & Schatz, M. C. (2018). Accurate detection of complex structural variations using single-molecule sequencing. *Nature Methods*, *15*(6), 461-+. https://doi.org/10.1038/s41592-018-0001-7
Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, *30*(14), 2068-2069. https://doi.org/10.1093/bioinformatics/btu153
Seemann, T. (2015). Snippy: fast bacterial variant calling from NGS reads.
Seemann, T. (2017). Shovill: Faster SPAdes assembly of Illumina reads. GitHub. In S. Nair (Ed.).
Sekyere, J. O., & Reta, M. A. (2020). Genomic and Resistance Epidemiology of Gram-Negative Bacteria in Africa: a Systematic Review and Phylogenomic Analyses from a One Health Perspective. *Msystems*, *5*(6). https://doi.org/ARTN e00897-20
10.1128/mSystems.00897-20
Sheppard, S. K., Dallas, J. F., Strachan, N. J., MacRae, M., McCarthy, N. D., Wilson, D. J., Gormley, F. J., Falush, D., Ogden, I. D., & Maiden, M. C. (2009). Campylobacter genotyping to determine the source of human infection. *Clinical Infectious Diseases*, *48*(8), 1072-1078.
Sheppard, S. K., Didelot, X., Jolley, K. A., Darling, A. E., Pascoe, B., Meric, G., Kelly, D. J., Cody, A., Colles, F. M., & Strachan, N. J. (2013). Progressive genome-wide introgression in agricultural Campylobacter coli. *Molecular ecology*, *22*(4), 1051-1064.
Sherry, N. L., Horan, K. A., Ballard, S. A., da Silva, A. G., Gorrie, C. L., Schultz, M. B., Stevens, K., Valcanis, M., Sait, M. L., Stinear, T. P., Howden, B. P., & Seemann, T. (2023). An ISO-certified genomics workflow for identification and surveillance of antimicrobial resistance. *Nature Communications*, *14*(1). https://doi.org/ARTN 60
10.1038/s41467-022-35713-4
Sia, C. M., Pearson, J. S., Howden, B. P., Williamson, D. A., & Ingle, D. J. (2025). Salmonella pathogenicity islands in the genomic era. *Trends Microbiol*. https://doi.org/10.1016/j.tim.2025.02.007
Song, S., Lee, B., Yeom, J.-H., Hwang, S., Kang, I., Cho, J.-C., Ha, N.-C., Bae, J., Lee, K., & Kim, Y.-H. (2015). MdsABC-mediated pathway for pathogenicity in Salmonella enterica serovar Typhimurium. *Infection and immunity*, *83*(11), 4266-4276.

Srirungruang, S., Mahajindawong, B., Nimitpanya, P., Bunkasem, U., Ayuyoe, P., Nuchprayoon, S., & Sanprasert, V. (2022). Comparative Study of DNA Extraction Methods for the PCR Detection of Intestinal Parasites in Human Stool Samples. *Diagnostics*, *12*(11). https://doi.org/ARTN 2588 10.3390/diagnostics12112588

Stadler, T., Meinel, D., Aguilar-Bultet, L., Huisman, J. S., Schindler, R., Egli, A., Seth-Smith, H. M. B., Eichenberger, L., Brodmann, P., Hübner, P., Bagutti, C., & Tschudin-Sutter, S. (2018). Transmission of ESBL-producing Enterobacteriaceae and their mobile genetic elements-identification of sources by whole genome sequencing: study protocol for an observational study in Switzerland. *Bmj Open*, *8*(2). https://doi.org/ARTN e021823 10.1136/bmjopen-2018-021823

Stinson, L. F., Keelan, J. A., & Payne, M. S. (2019). Profiling bacterial communities in low biomass samples: pitfalls and considerations. *Microbiology Australia*, *40*(4), 181-185.

Strepis, N., Dollee, D., Vrins, D., Vanneste, K., Bogaerts, B., Carrillo, C., Bharat, A., Horan, K., Sherry, N. L., Seemann, T., Howden, B. P., Hiltemann, S., Chindelevitch, L., Stubbs, A. P., & Hays, J. P. (2025). BenchAMRking: a Galaxy-based platform for illustrating the major issues associated with current antimicrobial resistance (AMR) gene prediction workflows. *Bmc Genomics*, *26*(1). https://doi.org/ARTN 27 10.1186/s12864-024-11158-5

Struelens, M. J., Ludden, C., Werner, G., Sintchenko, V., Jokelainen, P., & Ip, M. (2024). Real-time genomic surveillance for enhanced control of infectious diseases and antimicrobial resistance. *Frontiers in Science*, *2*, 1298248.

Takaya, A., Tomoyasu, T., Matsui, H., & Yamamoto, T. (2004). The DnaK/DnaJ chaperone machinery of Salmonella enterica serovar Typhimurium is essential for invasion of epithelial cells and survival within macrophages, leading to systemic infection. *Infection and immunity*, *72*(3), 1364-1373.

Tang, S. L., Orsi, R. H., Luo, H., Ge, C. T., Zhang, G. T., Baker, R. C., Stevenson, A., & Wiedmann, M. (2019). Assessment and Comparison of Molecular Subtyping and Characterization Methods for. *Frontiers in Microbiology*, *10*. https://doi.org/ARTN 1591 10.3389/fmicb.2019.01591

Tatusova, T., DiCuccio, M., Badretdin, A., Chetvernin, V., Nawrocki, E. P., Zaslavsky, L., Lomsadze, A., Pruitt, K. D., Borodovsky, M., & Ostell, J. (2016). NCBI prokaryotic genome annotation pipeline. *Nucleic Acids Res*, *44*(14), 6614-6624. https://doi.org/10.1093/nar/gkw569

Taylor, T. L., Volkening, J. D., DeJesus, E., Simmons, M., Dimitrov, K. M., Tillman, G. E., Suarez, D. L., & Afonso, C. L. (2019). Rapid, multiplexed, whole genome and plasmid sequencing of foodborne pathogens using long-read nanopore technology. *Scientific reports*, *9*. https://doi.org/ARTN 16350 10.1038/s41598-019-52424-x

Thépault, A., Méric, G., Rivoal, K., Pascoe, B., Mageiros, L., Touzain, F., Rose, V., Béven, V., Chemaly, M., & Sheppard, S. K. (2017). Genome-Wide Identification of Host-Segregating Epidemiological Markers for Source Attribution in. *Applied and Environmental Microbiology*, *83*(7). https://doi.org/ARTN e03085
10.1128/AEM.03085-16
Thystrup, C., Brinch, M. L., Henri, C., Mughini-Gras, L., Franz, E., Wieczorek, K., Gutierrez, M., Prendergast, D. M., Duffy, G., Burgess, C. M., Bolton, D., Alvarez, J., Lopez-Chavarrias, V., Rosendal, T., Clemente, L., Amaro, A., Zomer, A. L., Joensen, K. G., Nielsen, E. M., . . . Hald, T. (2025). Source attribution of human *Campylobacter* infection: a multi-country model in the European Union. *Frontiers in Microbiology*, *16*. https://doi.org/ARTN 1519189
10.3389/fmicb.2025.1519189
Toboldt, A., Tietze, E., Helmuth, R., Junker, E., Fruth, A., & Malorny, B. (2013). Population structure of Salmonella enterica serovar 4,[5], 12: b:− strains and likely sources of human infection. *Applied and Environmental Microbiology*, *79*(17), 5121-5129.
Tonkin-Hill, G., Ling, C., Chaguza, C., Salter, S. J., Hinfonthong, P., Nikolaou, E., Tate, N., Pastusiak, A., Turner, C., & Chewapreecha, C. (2022). Pneumococcal within-host diversity during colonization, transmission and treatment. *Nature Microbiology*, *7*(11), 1791-1804.
Treangen, T. J., & Salzberg, S. L. (2012). Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature reviews genetics*, *13*(1), 36-46.
Trigodet, F., Lolans, K., Fogarty, E., Shaiber, A., Morrison, H. G., Barreiro, L., Jabri, B., & Eren, A. M. (2022). High molecular weight DNA extraction strategies for long-read sequencing of complex metagenomes. *Molecular ecology resources*, *22*(5), 1786-1802. https://doi.org/10.1111/1755-0998.13588
Trivett, H., Darby, A. C., & Oyebode, O. (2025). Academic and clinical perspectives of metagenome sequencing as a diagnostic tool for infectious disease: an interpretive phenomenological study. *BMC Infectious Diseases*, *25*(1), 448.
Urban, L., Perlas, A., Francino, O., Martí-Carreras, J., Muga, B. A., Mwangi, J. W., Boykin Okalebo, L., Stanton, J. A. L., Black, A., & Waipara, N. (2023). Real-time genomics for One Health. *Molecular systems biology*, *19*(8), e11686.
Uritskiy, G. V., DiRuggiero, J., & Taylor, J. (2018). MetaWRAP-a flexible pipeline for genome-resolved metagenomic data analysis. *Microbiome*, *6*(1), 158. https://doi.org/10.1186/s40168-018-0541-1
Vasala, A., Hytönen, V. P., & Laitinen, O. H. (2020). Modern Tools for Rapid Diagnostics of Antimicrobial Resistance. *Frontiers in Cellular and Infection Microbiology*, *10*. https://doi.org/ARTN 308
10.3389/fcimb.2020.00308

Vicedomini, R., Quince, C., Darling, A. E., & Chikhi, R. (2021). Strainberry: automated strain separation in low-complexity metagenomes using long reads. *Nature Communications*, *12*(1), 4485.

Waldram, A., Dolan, G., Ashton, P. M., Jenkins, C., & Dallman, T. J. (2018). Epidemiological analysis of *Salmonella* clusters identified by whole genome sequencing, England and Wales 2014. *Food Microbiology*, *71*, 39-45. https://doi.org/10.1016/j.fm.2017.02.012

Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C. A., Zeng, Q., Wortman, J., & Young, S. K. (2014). Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *Plos One*, *9*(11), e112963.

Wang, X. Y., Zhu, S. L., Zhao, J. H., Bao, H. X., Liu, H. D., Ding, T. M., Liu, G. R., Li, Y. G., Johnston, R. N., Cao, F. L., Tang, L., & Liu, S. L. (2019). Genetic boundaries delineate the potential human pathogen

into discrete lineages: divergence and speciation. *Bmc Genomics*, *20*(1). https://doi.org/ARTN 930

10.1186/s12864-019-6259-z

Warburton, P. E., & Sebra, R. P. (2023). Long-Read DNA Sequencing: Recent Advances and Remaining Challenges. *Annual Review of Genomics and Human Genetics*, *24*, 109-132. https://doi.org/10.1146/annurev-genom-101722-103045

Warsi, O. M., Andersson, D. I., & Dykhuizen, D. E. (2018). Different adaptive strategies in E. coli populations evolving under macronutrient limitation and metal ion limitation. *BMC Evolutionary Biology*, *18*, 1-15.

Wasfy, M., Oyofo, B., Elgindy, A., & Churilla, A. (1995). Comparison of Preservation Media for Storage of Stool Samples. *Journal of Clinical Microbiology*, *33*(8), 2176-2178. https://doi.org/Doi 10.1128/Jcm.33.8.2176-2178.1995

Waters, E. V., Cameron, S. K., Langridge, G. C., & Preston, A. (2025). Bacterial genome structural variation: prevalence, mechanisms, and consequences. *Trends in Microbiology*.

Waters, E. V., Lee, W. W., Ismail Ahmed, A., Chattaway, M.-A., & Langridge, G. C. (2024). From acute to persistent infection: revealing phylogenomic variations in Salmonella Agona. *PLoS pathogens*, *20*(10), e1012679.

Waters, E. V., Tucker, L. A., Ahmed, J. K., Wain, J., & Langridge, G. C. (2022). Impact of Salmonella genome rearrangement on gene expression. *Evolution Letters*, *6*(6), 426-437. https://doi.org/10.1002/evl3.305

Wegl, G., Grabner, N., Köstelbauer, A., Klose, V., & Ghanbari, M. (2021). Toward Best Practice in Livestock Microbiota Research: A Comprehensive Comparison of Sample Storage and DNA Extraction Strategies. *Frontiers in Microbiology*, *12*. https://doi.org/ARTN 627539

10.3389/fmicb.2021.627539

WHO. (2022). WHO implementation handbook for national action plans on antimicrobial resistance: guidance for the human health sector. In *WHO implementation handbook for national action plans on antimicrobial resistance: guidance for the human health sector*.

Wick, R., & Menzel, P. (2019). Filtlong: quality filtering tool for long reads. *Github.[Google Scholar]*.

Wick, R. R., & Holt, K. E. (2022). Polypolish: Short-read polishing of long-read bacterial genome assemblies. *Plos Computational Biology*, *18*(1). https://doi.org/ARTN e1009802
10.1371/journal.pcbi.1009802

Wick, R. R., Judd, L. M., Gorrie, C. L., & Holt, K. E. (2017a). Completing bacterial genome assemblies with multiplex MinION sequencing. *Microbial Genomics*, *3*(10). https://doi.org/ARTN 000132
10.1099/mgen.0.000132

Wick, R. R., Judd, L. M., Gorrie, C. L., & Holt, K. E. (2017b). Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *Plos Computational Biology*, *13*(6), e1005595. https://doi.org/10.1371/journal.pcbi.1005595

Wick, R. R., Judd, L. M., & Holt, K. E. (2023). Assembling the perfect bacterial genome using Oxford Nanopore and Illumina sequencing. *Plos Computational Biology*, *19*(3). https://doi.org/ARTN e1010905
10.1371/journal.pcbi.1010905

Wilkinson, D. J., Dickins, B., Robinson, K., & Winter, J. A. (2022). Genomic diversity of *Helicobacter pylori* populations from different regions of the human stomach. *Gut microbes*, *14*(1), 2152306.

Wood, D. E., Lu, J., & Langmead, B. (2019). Improved metagenomic analysis with Kraken 2. *Genome Biology*, *20*(1), 257. https://doi.org/10.1186/s13059-019-1891-0

Woodcock, D. J., Krusche, P., Strachan, N. J., Forbes, K. J., Cohan, F. M., Méric, G., & Sheppard, S. K. (2017). Genomic plasticity and rapid host switching can promote the evolution of generalism: a case study in the zoonotic pathogen *Campylobacter*. *Scientific reports*, *7*(1), 9650.

Wright, M. H., Adelskov, J., & Greene, A. C. (2017). Bacterial DNA Extraction Using Individual Enzymes and Phenol/Chloroform Separation. *Journal of Microbiology & Biology Education*, *18*(2). https://doi.org/10.1128/jmbe.v18i2.1348

Wu, R. C., Payne, M., Zhang, L., & Lan, R. T. (2024). Uncovering the boundaries of *Campylobacter* species through large-scale phylogenetic and nucleotide identity analyses. *Msystems*, *9*(4). https://doi.org/ARTN e01218-23
10.1128/msystems.01218-23

Wylezich, C., Papa, A., Beer, M., & Höper, D. (2018). A Versatile Sample Processing Workflow for Metagenomic Pathogen Detection. *Scientific reports*, *8*. https://doi.org/ARTN 13108
10.1038/s41598-018-31496-1

Yamamoto, S., Iyoda, S., & Ohnishi, M. (2021). Stabilizing Genetically Unstable Simple Sequence Repeats in the Campylobacter jejuni Genome by Multiplex Genome Editing: a Reliable Approach for Delineating Multiple Phase-Variable Genes. *mBio*, *12*(4), e0140121. https://doi.org/10.1128/mBio.01401-21

Yang, F. M., Sun, J. H., Luo, H. N., Ren, H. H., Zhou, H. C., Lin, Y. X., Han, M., Chen, B., Liao, H. L., Brix, S., Li, J. H., Yang, H. M., Kristiansen, K., & Zhong, H. Z. (2020). Assessment of fecal DNA extraction protocols for metagenomic studies. *Gigascience*, *9*(7). https://doi.org/ARTN giaa071 10.1093/gigascience/giaa071

Yek, C., Pacheco, A. R., Vanaerschot, M., Bohl, J. A., Fahsbender, E., Aranda-Díaz, A., Lay, S., Chea, S., Oum, M. H., & Lon, C. (2022). Metagenomic pathogen sequencing in resource-scarce settings: lessons learned and the road ahead. *Frontiers in epidemiology*, *2*, 926695.

Yue, M., Rankin, S. C., Blanchet, R. T., Nulton, J. D., Edwards, R. A., & Schifferli, D. M. (2012). Diversification of the *Salmonella* Fimbriae: A Model of Macro- and Microevolution. *Plos One*, *7*(6). https://doi.org/ARTN e38596 10.1371/journal.pone.0038596

Zankari, E., Allesøe, R., Joensen, K. G., Cavaco, L. M., Lund, O., & Aarestrup, F. M. (2017). PointFinder: a novel web tool for WGS-based detection of antimicrobial resistance associated with chromosomal point mutations in bacterial pathogens. *Journal of antimicrobial chemotherapy*, *72*(10), 2764-2768.

Zankari, E., Hasman, H., Cosentino, S., Vestergaard, M., Rasmussen, S., Lund, O., Aarestrup, F. M., & Larsen, M. V. (2012). Identification of acquired antimicrobial resistance genes. *Journal of antimicrobial chemotherapy*, *67*(11), 2640-2644.

Zankari, E., Hasman, H., Kaas, R. S., Seyfarth, A. M., Agerso, Y., Lund, O., Larsen, M. V., & Aarestrup, F. M. (2013). Genotyping using whole-genome sequencing is a realistic alternative to surveillance based on phenotypic antimicrobial susceptibility testing. *Journal of antimicrobial chemotherapy*, *68*(4), 771-777. https://doi.org/10.1093/jac/dks496

Zarske, M., Luu, H. Q., Deneke, C., Knüver, M. T., Thieck, M., Hoang, H. T., Bretschneider, N., Pham, N. T., Huber, I., & Stingl, K. (2024). Identification of knowledge gaps in whole-genome sequence analysis of multi-resistant thermotolerant *Campylobacter* spp. *Bmc Genomics*, *25*(1). https://doi.org/ARTN 156 10.1186/s12864-024-10014-w

Zhang, H., Jain, C., & Aluru, S. (2020). A comprehensive evaluation of long read error correction methods. *Bmc Genomics*, *21*, 1-15.

Zhang, S. K., den Bakker, H. C., Li, S. T., Chen, J., Dinsmore, B. A., Lane, C., Lauer, A. C., Fields, P. I., & Deng, Y. (2019). SeqSero2: Rapid and Improved Serotype Determination Using Whole-Genome Sequencing Data. *Appl Environ Microbiol*, *85*(23). https://doi.org/10.1128/AEM.01746-19

Zhao, S., Tyson, G. H., Chen, Y., Li, C., Mukherjee, S., Young, S., Lam, C., Folster, J. P., Whichard, J. M., & McDermott, P. F. (2016). Whole-genome sequencing analysis accurately predicts antimicrobial resistance phenotypes in *Campylobacter* spp. *Applied and Environmental Microbiology*, *82*(2), 459-466. https://doi.org/10.1128/Aem.02873-15

Zhao, W. X., Zeng, W., Pang, B., Luo, M., Peng, Y., Xu, J. L., Kan, B., Li, Z. P., & Lu, X. (2023). Oxford nanopore long-read sequencing enables the

generation of complete bacterial and plasmid genomes without short-read sequencing. *Frontiers in Microbiology*, *14*. https://doi.org/ARTN 1179966
10.3389/fmicb.2023.1179966

Zhou, Z. M., Alikhan, N. F., Mohamed, K., Fan, Y. L., Achtman, M., Brown, D., Chattaway, M., Dallman, T., Delahay, R., Kornschober, C., Pietzka, A., Malorny, B., Petrovska, L., Davies, R., Robertson, A., Tyne, W., Weill, F. X., Accou-Demartin, M., Williams, N., & Grp, A. S. (2020). The EnteroBase user's guide, with case studies on *Salmonella* transmissions, *Yersinia pestis* phylogeny, and *Escherichia* core genomic diversity. *Genome Research*, *30*(1), 138-152. https://doi.org/10.1101/gr.251678.119

Zidane, N., Rodrigues, C., Bouchez, V., Rethoret-Pasty, M., Passet, V., Brisse, S., & Crestani, C. (2025). Accurate genotyping of three major respiratory bacterial pathogens with ONT R10. 4.1 long-read sequencing. *Genome Research*, gr. 279829.279124.

Appendix 1 - Key Developmental Protocol Variants

Fire Monkey base

To lyse the cells, 30 mg/µL lysozyme was added to a STET1 buffer containing 1.2% Triton X-100, 100 µL of this lysis buffer was added to the pellet. Samples were pipette-mixed 5 times and briefly vortexed (10 seconds) before incubating at 37°C for 10 minutes. A master mix of 300 µL LSDNA buffer and 20 µL Proteinase K was prepared for the appropriate number of samples. The 320 µL LSDNA Proteinase K mix was added to samples before pipette-mixing 5 times and brief vortexing (10 seconds). These samples were then incubated at 37°C for 20 minutes. After incubation, 20 µL RNase A solution was added to the samples, which were then rested at room temperature for 5 minutes. A 350 µL volume of BS was added to the samples, which were mixed by vortexing (10 seconds). Finally, a 400 µL volume of 75% isopropanol was added to the samples, which were mixed by vortexing (10 seconds).

FM-W

To process the sample, 2 mL of 50°C sterile water was added to the pellet. After adding the water, vortexing was performed for 30 seconds to ensure thorough mixing. The tube was then centrifuged at 18,000 rcf for 3 minutes to pellet the HD sample. Carefully, the supernatant was gently removed from the tube. The pellet was resuspended in another 2 mL of 50°C dH_2O, followed by centrifugation at 18,000 rcf for 3 minutes to pellet the HD sample again. The supernatant was gently removed once more. This resuspension process was repeated once more: the pellet was resuspended in 2 mL of 50°C dH_2O, followed by centrifugation at 18,000 rcf for 3 minutes, and gentle removal of the supernatant. To lyse the cells, 30 mg/µL lysozyme was added to a STET1 buffer, 100 µL of this lysis buffer was added to the pellet. Samples were pipette-mixed 5 times and briefly vortexed (10 seconds) before incubating at 37°C for 10 minutes. A master mix of 300 µL LSDNA buffer and 20 µL Proteinase K was prepared for the appropriate number of samples. The 320 µL LSDNA Proteinase K mix was added to samples before pipette mixing 5 times and brief vortexing (10 seconds). These samples were then incubated at 37°C for 20 minutes. After incubation, 20 µL RNase A solution was added to the samples, which were then rested at room temperature for 5 minutes. A 350 µL volume of BS was added to the samples, which were mixed by vortexing (10 seconds). Finally, a 400 µL volume

of 75% isopropanol was added to the samples, which were mixed by vortexing (10 seconds).

FM-W-3x

To process the sample, 2 mL of 50°C sterile water was added to the pellet. After adding the water, vortexing was performed for 30 seconds to ensure thorough mixing. The tube was then centrifuged at 18,000 rcf for 3 minutes to pellet the HD sample. Carefully, the supernatant was gently removed from the tube. The pellet was resuspended in another 2 mL of 50°C dH_2O, followed by centrifugation at 18,000 rcf for 3 minutes to pellet the HD sample again. The supernatant was gently removed once more. This resuspension process was repeated once more: the pellet was resuspended in 2 mL of 50°C dH_2O, followed by centrifugation at 18,000 rcf for 3 minutes, and gentle removal of the supernatant. A 2 mL tube was used for each sample. Initially, 300 µL of STET1 (30 mg/mL lysozyme) was added to the tube to facilitate lysis. The stool pellet was resuspended using a wide bore 1000 µL tip to ensure thorough mixing. Subsequently, a narrow bore 1000 µL tip was used for additional resuspension steps (x10). The mixture was then incubated at 37°C for 10 minutes. Following this, 900 µL of LSDNA and 60 µL of 20 mg/mL Proteinase K were added to the tube, and the contents were again resuspended using a wide bore tip and then a narrow bore tip (x10). The tube was incubated at 56°C for 20 minutes for further enzymatic digestion. Afterward, 3 µL of RNase A (100 ug/µL in H_2O) was added, and the tube was left at room temperature for 5 minutes. The lysate was split into three Eppendorf tubes, and each tube received 350 µL of BS, followed by resuspension using a narrow bore tip (x10). The tubes were then incubated at room temperature for 20 minutes. Subsequently, each tube was centrifuged at 18,000 x g for 20 minutes to pellet DNA. Carefully avoiding the yellow/brown oily solution at the tube bottom, the supernatant was transferred to fresh tubes. To precipitate DNA, 400 µL of 75% isopropanol was added to each tube, which was then vortexed. The samples were processed using an A200 plate column: tube 1 and tube 2 were run using the stool load only protocol, while tube 3 was processed using the full stool protocol. Elution was set for 1 x 100 µL, with the second elution yield being very low, prompting cessation of further collection. Clean the resulting DNA with SPRI beads with 0.6x-1x SPRI depending on how you want to size select your DNA fragments. During the SPRI bead

clean it helps if you resuspend the beads in ethanol off the magnet rather than just pipetting the ethanol over the beads while on the magnet.

Appendix 2 - Sequencing stats for metagenome samples used in Chapter 4. Stats include raw data yield, reads_in represents total read yield and reads_out represents the number of reads after human read removal.

Stool_id	Condition/Timepoint	Company	Raw_data	reads_in	reads_out	reads removed (%)
124	F0	Novogene	8.40	56066766	55802724	0.00471
124	G1	Azenta	16.52	110150344	110050194	0.00091
124	R1	Novogene	10.20	48447890	48360118	0.00181
124	Z1	Novogene	9.80	65028558	64747528	0.00432
124	G3	Novogene	13.73	91530664	91370456	0.00175
124	R3	Novogene	13.42	89484174	89417182	0.00075
124	Z3	Novogene	12.79	85291818	84896796	0.00463
124	G9	Novogene	12.26	81741296	81566394	0.00214
124	R9	Novogene	15.10	100687760	100605266	0.00082
124	Z9	Novogene	10.33	68851612	68607974	0.00354
130_R1	F0	Novogene	12.50	65560544	56974548	0.13096
130_R1	G1	Novogene	8.09	53908186	48300956	0.10401
130_R1	R1	Novogene	8.16	54389832	47908504	0.11916
130_R1	Z1	Novogene	9.36	62429170	17423208	0.72091
130_R1	G3	Novogene	11.79	78601722	67842056	0.13689
130_R1	R3	Azenta	9.80	65334054	59102272	0.09538
130_R1	Z3	Novogene	7.75	51685494	14307996	0.72317
130_R1	G9	Novogene	10.49	69932940	63696428	0.08918
130_R1	R9	Novogene	11.26	75047520	68104596	0.09251
130_R1	Z9	Novogene	9.93	66225976	18465254	0.72118
130_R2	F0	Novogene	12.70	78231396	65713446	0.16001
130_R2	G1	Azenta	13.99	93254392	85363352	0.08462
130_R2	R1	Azenta	13.48	89860284	80729094	0.10162
130_R2	Z1	Azenta	12.48	83174914	22804414	0.72583
130_R2	G3	Novagene	9.59	63920536	53075422	0.16967
130_R2	R3	Novagene	10.36	51685494	14307996	0.72317
130_R2	Z3	Azenta	11.07	73800650	21760322	0.70515
130_R2	G9	Novagene	10.04	66947236	61287338	0.08454
130_R2	R9	Novagene	12.76	85098918	74924612	0.11956
130_R2	Z9	Novagene	8.28	66225976	18465254	0.72118
132	F0	Novogene	12.30	82103488	80372124	0.02109
132	G1	Novogene	10.50	69999080	68997446	0.01431
132	R1	Novogene	8.87	59103578	58822428	0.00476
132	Z1	Azenta	8.80	58681304	55807330	0.04898
132	G3	Novogene	11.28	75210316	73847238	0.01812
132	R3	Novogene	10.88	72521542	71979968	0.00747
132	Z3	Novogene	9.55	63681826	55229530	0.13273

Appendix 2 - Sequencing stats for metagenome samples used in Chapter 4. Stats include raw data yield, reads_in represents total read yield and reads_out represents the number of reads after human read removal.

Stool_id	Condition/Timepoint	Company	Raw_data	reads_in	reads_out	reads removed (%)
132	G9	Novogene	10.21	68089438	67231656	0.0126
132	R9	Novogene	11.16	74380490	73856578	0.00704
132	Z9	Azenta	10.55	70335360	66868874	0.04929
135_R1	F0	Novogene	9.8	83099226	82897604	0.00243
135_R1	G1	Novogene	10.13	67527456	67358820	0.0025
135_R1	R1	Novogene	8.6	57312146	57145750	0.0029
135_R1	Z1	Azenta	15.07	100468290	100396976	0.00071
135_R1	G3	Novogene	10.32	68813958	68690088	0.0018
135_R1	R3	Novogene	13.29	88619282	88354654	0.00299
135_R1	Z3	Azenta	9.23	61541608	60836538	0.01146
135_R1	G9	Novogene	11.44	76277592	76183842	0.00123
135_R1	R9	Novogene	13.07	87105596	87038918	0.00077
135_R1	Z9	Novogene	8.55	56988826	55834812	0.02025
135_R2	F0	Novogene	11.7	84836186	84667994	0.00198
135_R2	G1	Novogene	9.22	61460382	61319136	0.0023
135_R2	R1	Novogene	9.96	66398304	66200950	0.00297
135_R2	Z1	Novogene	8.43	56227236	55460722	0.01363
135_R2	G3	Novogene	10.23	68214230	68027292	0.00274
135_R2	R3	Novogene	11.15	74348210	74144668	0.00274
135_R2	Z3	Azenta	9.86	65725016	65147778	0.00878
135_R2	G9	Novogene	10.51	70058002	69933746	0.00177
135_R2	R9	Novogene	12.14	80963396	80882664	0.001
135_R2	Z9	Azenta	9.9	65980016	65402784	0.00875
136	F0	Novogene	11.36	75702524	74794248	0.012
136	G1	Novogene	10.38	69201452	68370348	0.01201
136	R1	Novogene	11.41	76040940	75511344	0.00696
136	Z1	Novogene	9.86	65752294	64784764	0.01471
136	G3	Novogene	9.07	60487756	59734616	0.01245
136	R3	Novogene	11.5	76671074	76021232	0.00848
136	Z3	Novogene	9.44	62930336	62161734	0.01221
136	G9	Novogene	8.76	58392242	57672010	0.01233
136	R9	Novogene	11.95	79650832	79059108	0.00743
136	Z9	Azenta	12.54	83596152	83219096	0.00451
141	F0	Novogene	10.8	72019188	38843240	0.46065
141	G1	Novogene	9.97	66464860	5328346	0.91983
141	R1	Novogene	9.11	60715036	4244108	0.9301
141	Z1	Azenta	18.32	122156238	43599168	0.64309

Appendix 2 - Sequencing stats for metagenome samples used in Chapter 4. Stats include raw data yield, reads_in represents total read yield and reads_out represents the number of reads after human read removal.

Stool_id	Condition/Timepoint	Company	Raw_data	reads_in	reads_out	reads removed (%)
141	G3	Novogene	10.19	67956122	7599462	0.88817
141	R3	Novogene	9.41	62754532	3899346	0.93786
141	Z3	Azenta	12.74	84939984	4505360	0.94696
141	G9	Azenta	4.97	33139694	1834604	0.94464
141	R9	Azenta	10.49	69914624	40164084	0.42553
141	Z9	Azenta	7.74	51603250	1983408	0.96156
143	F0	Novogene	14.36	95747506	9993000	0.89563
143	G1	Novogene	10.76	71738480	67860998	0.05405
143	R1	Novogene	9.47	63123554	59690696	0.05438
143	Z1	Novogene	7.99	53263636	4157512	0.92194
143	G3	Novogene	8.15	54319056	51844870	0.04555
143	R3	Novogene	9.33	62222744	2988864	0.95197
143	Z3	Azenta	13.91	92744048	88584236	0.04485
143	G9	Azenta	9.79	65241864	62892766	0.03601
143	R9	Azenta	9.68	64561976	62474772	0.03233
143	Z9	Azenta	11.03	73541766	4209340	0.94276
144	F0	Novogene	9.04	60288958	59534980	0.01251
144	G1	Novogene	12.27	81772644	81196398	0.00705
144	R1	Novogene	8.4	56021382	55694316	0.00584
144	Z1	Novogene	10.3	68697528	60278126	0.12256
144	G3	Azenta	12.04	80289382	79679610	0.00759
144	R3	Azenta	10.99	73245362	72811348	0.00593
144	Z3	Azenta	10.11	67374118	64107740	0.04848
144	G9	Azenta	10.95	73004802	72564852	0.00603
144	R9	Azenta	9	59999222	59807580	0.00319
144	Z9	Azenta	10.59	70616660	68204108	0.03416
145	F0	Novogene	8.1	53999702	53841384	0.00293
145	G1	Novogene	8.28	72603804	69465536	0.04322
145	R1	Novogene	9.14	73871914	68802074	0.06863
145	Z1	Novogene	8.63	57553284	56994950	0.0097
145	G3	Azenta	10.29	68615464	68452726	0.00237
145	R3	Azenta	11.28	75194812	74988478	0.00274
145	Z3	Azenta	10.21	68065130	67520526	0.008
145	G9	Azenta	11.78	78559188	78388024	0.00218
145	R9	Azenta	12.31	82092636	81806014	0.00349
145	Z9	Azenta	10.79	71914528	71539282	0.00522
146	F0	Novogene	10.25	68315170	31091866	0.54488

Appendix 2 - Sequencing stats for metagenome samples used in Chapter 4. Stats include raw data yield, reads_in represents total read yield and reads_out represents the number of reads after human read removal.

Stool_id	Condition/Timepoint	Company	Raw_data	reads_in	reads_out	reads removed (%)
146	G1	Novogene	10.89	72603804	69465536	0.04322
146	R1	Novogene	11.08	73871914	68802074	0.06863
146	Z1	Azenta	9.05	60323412	17313812	0.71298
146	G3	Azenta	13.28	88507640	82256076	0.07063
146	R3	Azenta	11.71	78041878	63798140	0.18251
146	Z3	Azenta	10.7	71309860	17904176	0.74892

Appendix 3 - AMR in-silico predictions for isolates and MD-Campylobacter genomes used in Chapter 4

Isolate	Macrolide	Quinolone	Beta-lactamase (unknown spectrum)	Tetracycline
124-6	50S_L22_A103V	gyrA_T86I	blaOXA-193	tet(O)*
124-1	50S_L22_A103V	gyrA_T86I	blaOXA-193	tet(O)*
124-7	50S_L22_A103V	gyrA_T86I	blaOXA-193	tet(O)*
124-10	50S_L22_A103V	gyrA_T86I	blaOXA-193	tet(O)*
124-8	50S_L22_A103V	gyrA_T86I	blaOXA-193	tet(O)*
124-3	50S_L22_A103V	gyrA_T86I	blaOXA-193	tet(O)*
124-11	50S_L22_A103V	gyrA_T86I	blaOXA-193	tet(O)*
124-5	50S_L22_A103V	gyrA_T86I	blaOXA-193	tet(O)*
124-12	50S_L22_A103V	gyrA_T86I	blaOXA-193	tet(O)*
124-2	50S_L22_A103V	gyrA_T86I	blaOXA-193	tet(O)*
MDG	Macrolide	Quinolone	Beta-lactamase (unknown spectrum)	Tetracycline
124_TP0	50S_L22_A103V	gyrA_T86I	-	-
124_G1M	50S_L22_A103V	-	-	-
124_G3M	50S_L22_A103V	gyrA_T86I	blaOXA-193^	-
124_G9M	-	gyrA_T86I	-	-
124_R1M	-	-	-	-
124_R3M	50S_L22_A103V		-	-
124_R9M	-	-	-	-
124_Z1M	-	-	-	-
124_Z3M	50S_L22_A103V	-	blaOXA-193^	-
124_Z9M	50S_L22_A103V	-	blaOXA-193^	-

Appendix 3 - AMR in-silico predictions for isolates and MD-Campylobacter genomes used in Chapter 4

Isolate	Beta-lactamase (unknown spectrum)	Tetracycline
130-2	blaOXA-184	-
130-6	blaOXA-184	tet(O)*
130-3	blaOXA-184	tet(O)*
130-8	blaOXA-184	tet(O)*
130-4	blaOXA-184	tet(O)*
130-10	blaOXA-184	tet(O)*
130-7	blaOXA-184	tet(O)*
130-1	blaOXA-184	tet(O)*
130-9	blaOXA-184	tet(O)*
MDG	Beta-lactamase (unknown spectrum)	Tetracycline
130r1_TP0	-	-
130r1_G1M	-	-
130r1_G3M	-	-
130r1_G9M	-	-
130r1_R1M	-	-
130r1_R3M	-	-
130r1_R9M	-	-
130r1_Z1M	-	-
130r1_Z3M	-	-
130r1_Z9M	-	-
130r2_TP0	-	-
130r2_G1M	-	-
130r2_G3M	-	-
130r2_G9M	-	-

Appendix 3 - AMR in-silico predictions for isolates and MD-Campylobacter genomes used in Chapter 4

MDG	Beta-lactamase (unknown spectrum)	Tetracycline
130r2_R1M	-	-
130r2_R3M	-	-
130r2_R9M	-	-
130r2_Z1M	-	-
130r2_Z3M	-	-
130r2_Z9M	-	-

Appendix 3 - AMR in-silico predictions for isolates and MD-Campylobacter genomes used in Chapter 4

Isolate	Macrolide	Quinolone	Beta-lactamase (unknown spectrum)	Tetracycline
132C-8	50S_L22_A103V	gyrA_T86I	blaOXA-193	tet(O)*
132C-11	50S_L22_A103V	gyrA_T86I	blaOXA-193	tet(O)*
132C-3	50S_L22_A103V	gyrA_T86I	blaOXA-193	tet(O)*
132C-1	50S_L22_A103V	gyrA_T86I	blaOXA-193	tet(O)*
132C-7	50S_L22_A103V	gyrA_T86I	blaOXA-193	tet(O)*
132C-9	50S_L22_A103V	gyrA_T86I	blaOXA-193	tet(O)*
132C-6	50S_L22_A103V	gyrA_T86I	blaOXA-193	tet(O)*
132C-10	50S_L22_A103V	gyrA_T86I	blaOXA-193	tet(O)*
132C-2	50S_L22_A103V	gyrA_T86I	blaOXA-193	tet(O)*
132C-4	50S_L22_A103V	gyrA_T86I	blaOXA-193	tet(O)*
132C-5	50S_L22_A103V	gyrA_T86I	blaOXA-193	tet(O)*
132C-12	50S_L22_A103V	gyrA_T86I	blaOXA-193	tet(O)*
MDG	Macrolide	Quinolone	Beta-lactamase (unknown spectrum)	Tetracycline
132_TP0.fasta	50S_L22_A103V	gyrA_T86I	blaOXA-193	
132_G_1M.fasta	50S_L22_A103V	gyrA_T86I	blaOXA-193	
132_G_3M.fasta	50S_L22_A103V	gyrA_T86I	blaOXA-193	
132_G_9M.fasta	50S_L22_A103V	gyrA_T86I	blaOXA-193	
132_R_1M.fasta	50S_L22_A103V	gyrA_T86I	blaOXA-193	
132_R_3M.fasta	50S_L22_A103V	gyrA_T86I	blaOXA-193^	
132_R_9M.fasta	50S_L22_A103V	gyrA_T86I	blaOXA-193	
132_Z_1M.fasta	50S_L22_A103V	gyrA_T86I	blaOXA-193	
132_Z_3M.fasta	50S_L22_A103V	gyrA_T86I	blaOXA-193	
132_Z_9M.fasta	50S_L22_A103V	gyrA_T86I	blaOXA-193	

Appendix 3 - AMR in-silico predictions for isolates and MD-Campylobacter genomes used in Chapter 4

Isolate	Quinolone	Beta-lactamase (unknown spectrum)	Tetracycline
135-6	gyrA_T86I	blaOXA-193	tet(O)*
135-7	gyrA_T86I	blaOXA-193	tet(O)*
135-1	gyrA_T86I	blaOXA-193	tet(O)*
135-3	gyrA_T86I	blaOXA-193	tet(O)*
135-9	gyrA_T86I	blaOXA-193	tet(O)*
135-2	gyrA_T86I	blaOXA-193	tet(O)*
135-8	gyrA_T86I	blaOXA-193	tet(O)*
135-4	gyrA_T86I	blaOXA-193	tet(O)*
135-10	gyrA_T86I	blaOXA-193	tet(O)*
135-11	gyrA_T86I	blaOXA-193	tet(O)*
135-5	gyrA_T86I	blaOXA-193	tet(O)*
135-12	gyrA_T86I	blaOXA-193	tet(O)*
MDG	Quinolone	Beta-lactamase (unknown spectrum)	Tetracycline
135r1_TP0	gyrA_T86I	blaOXA-193	-
135r1_G1M	gyrA_T86I	blaOXA-193	-
135r1_G3M	gyrA_T86I	blaOXA-193	-
135r1_G9M	gyrA_T86I	blaOXA-193	-
135r1_R1M	gyrA_T86I	blaOXA-193	-
135r1_R3M	gyrA_T86I	blaOXA-193^	-
135r1_R9M	gyrA_T86I	blaOXA-193	-
135r1_Z1M	gyrA_T86I	blaOXA-193*	-
135r1_Z3M	-	blaOXA-193	-
135r1_Z9M	gyrA_T86I	blaOXA-193	-
135r2_TP0	gyrA_T86I	blaOXA-193	-

Appendix 3 - AMR in-silico predictions for isolates and MD-Campylobacter genomes used in Chapter 4

MDG	Quinolone	Beta-lactamase (unknown spectrum)	Tetracycline
135r2_G1M	gyrA_T86I	blaOXA-193	-
135r2_G3M	gyrA_T86I	blaOXA-193	-
135r2_G9M	gyrA_T86I	blaOXA-193	-
135r2_R1M	gyrA_T86I	blaOXA-193^	-
135r2_R3M	gyrA_T86I	blaOXA-193	-
135r2_R9M	gyrA_T86I	blaOXA-193	-
135r2_Z1M	gyrA_T86I	blaOXA-193	-
135r2_Z3M	gyrA_T86I	blaOXA-193	-
135r2_Z9M	gyrA_T86I	blaOXA-193	-

Appendix 3 - AMR in-silico predictions for isolates and MD-Campylobacter genomes used in Chapter 4

Isolate	Macrolide
136-6	50S_L22_A103V
136-8	50S_L22_A103V
136-12	50S_L22_A103V
136-4	50S_L22_A103V
136-2	50S_L22_A103V
136-10	50S_L22_A103V
MDG	Macrolide
136_TP0	50S_L22_A103V
136_G1M	50S_L22_A103V
136_G3M	50S_L22_A103V
136_G9M	50S_L22_A103V
136_R1M	50S_L22_A103V
136_R3M	50S_L22_A103V
136_R9M	50S_L22_A103V
136_Z1M	-
136_Z3M	-
136_Z9M	-

Appendix 3 - AMR in-silico predictions for isolates and MD-Campylobacter genomes used in Chapter 4

Isolate	Quinolone	Beta-lactamase (unknown spectrum)	Tetracycline
141-10	gyrA_T86I	blaOXA-193	tet(O)*
141-1	gyrA_T86I	blaOXA-193	tet(O)*
141-12	gyrA_T86I	blaOXA-193	tet(O)*
141-3	gyrA_T86I	blaOXA-193	tet(O)*
141-6	gyrA_T86I	blaOXA-193	tet(O)*
141-8	gyrA_T86I	blaOXA-193	tet(O)*
141-7	gyrA_T86I	blaOXA-193	tet(O)*
141-4	gyrA_T86I	blaOXA-193	tet(O)*
141-9	gyrA_T86I	blaOXA-193	tet(O)*
141-11	gyrA_T86I	blaOXA-193	tet(O)*
141-2	gyrA_T86I	blaOXA-193	tet(O)*
141-5	gyrA_T86I	blaOXA-193	tet(O)*
MDG	Quinolone	Beta-lactamase (unknown spectrum)	Tetracycline
141_TP0	gyrA_T86I	blaOXA-193*	-
141_G1M	-	-	-
141_G3M	-	-	-
141_G9M	-	-	-
141_R1M	-	blaOXA-193*	-
141_R3M	-	-	-
141_R9M	gyrA_T86I	blaOXA-193^	-
141_Z1M	-	-	-
141_Z3M	-	-	-
141_Z9M	-	-	-

Appendix 3 - AMR in-silico predictions for isolates and MD-Campylobacter genomes used in Chapter 4

Isolate	Quinolone	Beta-lactamase (unknown spectrum)
143-8	gyrA_T86I	blaOXA-193
143-3	gyrA_T86I	blaOXA-193
143-10	gyrA_T86I	blaOXA-193
143-4	gyrA_T86I	blaOXA-193
143-6	gyrA_T86I	blaOXA-193
143-5	gyrA_T86I	blaOXA-193
143-9	gyrA_T86I	blaOXA-193
143-11	gyrA_T86I	blaOXA-193
143-7	gyrA_T86I	blaOXA-193
143-12	gyrA_T86I	blaOXA-193
143-1	gyrA_T86I	blaOXA-193
143-2	gyrA_T86I	blaOXA-193
MDG	Quinolone	Beta-lactamase (unknown spectrum)
143_TP0	gyrA_T86I	blaOXA-193
143_G1M	gyrA_T86I	blaOXA-193
143_G_3M	gyrA_T86I	blaOXA-193
143_G_9M	gyrA_T86I	blaOXA-193
143_R1M	gyrA_T86I	blaOXA-193
143_R_3M	gyrA_T86I	blaOXA-193
143_R_9M	gyrA_T86I	blaOXA-193
143_Z_1M	-	-
143_Z3M	-	-
143_Z_9M	-	-

Appendix 3 - AMR in-silico predictions for isolates and MD-Campylobacter genomes used in Chapter 4

Isolate	Beta-lactamase (unknown spectrum)	Tetracycline	ESBL
144-12	blaOXA-193*	tet(L)*,tet(M)*,tet(O)*	-
144-8	blaOXA-193*	tet(L)*,tet(M)*,tet(O)*	-
144-6	blaOXA-193	tet(O)*	-
144-3	blaOXA-193	tet(O)*	-
144-1	blaOXA-193	tet(O)*	-
144-7	blaOXA-193	tet(L)*,tet(M)*,tet(O)*	-
144-11	blaOXA-193	tet(O)*	-
144-10	blaOXA-193	tet(O)*	-
144-4	blaOXA-193	tet(O)*	cepA
144-2	blaOXA-193	tet(O)*	-
144-5	blaOXA-193	tet(O)*	-
144-9	blaOXA-193	tet(L)*,tet(M)*,tet(O)*	-
MDG	Beta-lactamase (unknown spectrum)	Tetracycline	ESBL
144_TP0	-	-	-
144_G1M	-	-	-
144_G3M	-	-	-
144_G9M	-	-	-
144_R1M	-	-	-
144_R3M	-	-	-
144_R9M	-	-	-
144_Z1M	-	-	-
144_Z3M	-	-	-
144_Z9M	-	-	-

Appendix 3 - AMR in-silico predictions for isolates and MD-Campylobacter genomes used in Chapter 4

Isolate	Beta-lactamase (unknown spectrum)	Tetracycline	Quinolone	Beta-lactamase	Lincosamides	Trimetho-prim	Efflux	Strepto-mycin
145-7	blaOXA-489	tet(O)*,tet(Q)*	gyrA_T86I	cfxA*	-	-	-	-
145-4	blaOXA-489	tet(O)*,tet(Q)*	gyrA_T86I	cfxA*	-	-	-	-
145-11	blaOXA-489	tet(O)*,tet(Q)*	gyrA_T86I	cfxA*	-	-	-	-
145-12	blaOXA-489	tet(L)*,tet(O)*	gyrA_T86I	-	lnu(C)*	dfrF	-	-
145-8	blaOXA-489	tet(O)*	gyrA_T86I	cfxA*	-	-	-	-
145-5	blaOXA-489	tet(O)*,tet(Q)*	gyrA_T86I	cfxA*	-	-	-	-
145-6	blaOXA-489	tet(O)*,tet(X1)*,tet(X2)	gyrA_T86I	cfxA*	-	-	bexA*	aadS
145-3	blaOXA-489	tet(L)*,tet(O)*	gyrA_T86I	-	lnu(C)*	dfrF	-	-
145-1	blaOXA-489	tet(O)*	gyrA_T86I	-	-	-	-	-
145-10	blaOXA-489	tet(O)*	gyrA_T86I	cfxA*	-	-	-	-
145-2	blaOXA-489	tet(O)*,tet(X1)*,tet(X2)*	gyrA_T86I	cfxA*	-	-	bexA*	aadS
MDG	Beta-lactamase (unknown spectrum)	Tetracycline	Quinolone	Beta-lactamase	Lincosamides	Trimetho-prim	Efflux	Strepto-mycin
145_TP0	-	-	-	-	-	-	-	-
145_G1M	-	-	-	-	-	-	-	-
145_G3M	-	-	-	-	-	-	-	-
145_G9M	-	-	-	-	-	-	-	-
145_R1M	-	-	-	-	-	-	-	-
145_R3M	-	-	-	-	-	-	-	-
145_R9M	-	-	-	-	-	-	-	-
145_Z1M	-	-	-	-	-	-	-	-
145_Z3M	-	-	-	-	-	-	-	-
145_Z9M	-	-	-	-	-	-	-	-

Appendix 3 - AMR in-silico predictions for isolates and MD-Campylobacter genomes used in Chapter 4

Isolate	Beta-lactamase (unknown spectrum)
146-1	blaOXA-193
146-2	blaOXA-193
146-5	blaOXA-193
146-6	blaOXA-193
146-8	blaOXA-193
146-9	blaOXA-193
146-10	blaOXA-193
146-11	blaOXA-193
146-12	blaOXA-193
MDG	Beta-lactamase (unknown spectrum)
146_TP0	blaOXA-193^
146_G1M	blaOXA-193*
146_G3M	-
146_G9M	-
146_R1M	blaOXA-193^
146_R3M	-
146_R9M	-
146_Z1M	-
146_Z3M	-
146_Z9M	-

Appendix 3 - AMR in-silico predictions for isolates and MD-Campylobacter genomes used in Chapter 4

Isolate	Quinolone	Macrolide	Tetracycline
147-9	gyrA_T86I	50S_L22_A103V	tet(O)*
147-1	gyrA_T86I	50S_L22_A103V	tet(O)*
147-5	gyrA_T86I	50S_L22_A103V	tet(O)*
147-2	gyrA_T86I	50S_L22_A103V	tet(O)*
147-7	gyrA_T86I	50S_L22_A103V	tet(O)*
147-10	gyrA_T86I	50S_L22_A103V	tet(O)*
147-8	gyrA_T86I	50S_L22_A103V	tet(O)*
147-11	gyrA_T86I	50S_L22_A103V	tet(O)*
147-3	gyrA_T86I	50S_L22_A103V	tet(O)*
147-12	gyrA_T86I	50S_L22_A103V	tet(O)*
147-6	gyrA_T86I	50S_L22_A103V	tet(O)*
147-4	gyrA_T86I	50S_L22_A103V	tet(O)*
MDG	Quinolone	Macrolide	Tetracycline
147_TP0	gyrA_T86I	50S_L22_A103V	-
147_G1M	-	50S_L22_A103V	-
147_G3M	-	-	-
147_G9M	-	50S_L22_A103V	-
147_R1M	-	-	-
147_R3M	-	50S_L22_A103V	-
147_R9M	-	50S_L22_A103V	-
147_Z1M	-	-	-
147_Z3M	-	-	-
147_Z9M		-	-

Appendix 4 - Raw and normalised data input for statistical tests, MD-Campylobacter genomes used in Chapter 4

Stool_id	Conditions	reads_in	Breadth	Depth	Genome fraction	Breadth per 10M reads	Depth per 10M reads
124	F0	56066766	90.87	7.23	69.93	16.21	1.29
124	G1	110150344	64.69	2.3	5.25	5.87	0.21
124	R1	48447890	47.96	1.12	75.37	9.9	0.23
124	Z1	65028558	61.74	1.83	12.14	9.49	0.28
124	G3	91530664	94.14	6.3	12.33	10.29	0.69
124	R3	89484174	58.6	1.45	77.49	6.55	0.16
124	Z3	85291818	85.08	3.71	43.02	9.98	0.43
124	G9	81741296	88.97	4.45	9.18	10.88	0.54
124	R9	100687760	54.71	1.3	60.41	5.43	0.13
124	Z9	68851612	89.05	4.05	49.32	12.93	0.59
130_R1	F0	65560544	10.45	0.16	0	1.59	0.02
130_R1	G1	53908186	8.89	0.13	0	1.65	0.02
130_R1	R1	54389832	6.99	0.09	0	1.29	0.02
130_R1	Z1	62429170	0.99	0.01	0	0.16	0
130_R1	G3	78601722	12.49	0.18	0	1.59	0.02
130_R1	R3	65334054	8.61	0.13	0	1.32	0.02
130_R1	Z3	51685494	1.48	0.02	0	0.29	0
130_R1	G9	69932940	9.09	0.13	0.03	1.3	0.02
130_R1	R9	75047520	8.31	0.11	0	1.11	0.02
130_R1	Z9	66225976	0.94	0.01	0	0.14	0
130_R2	F0	78231396	10.01	0.15	0	1.28	0.02

Appendix 4 - Raw and normalised data input for statistical tests, MD-Campylobacter genomes used in Chapter 4

Stool_id	Conditions	reads_in	Breadth	Depth	Genome fraction	Breadth per 10M reads	Depth per 10M reads
130_R2	G1	93254392	11.66	0.22	0	1.25	0.02
130_R2	R1	89860284	10.06	0.18	0.03	1.12	0.02
130_R2	Z1	83174914	0.87	0.01	0	0.1	0
130_R2	G3	63920536	2.02	0.03	0	0.32	0.01
130_R2	R3	51685494	0.82	0.01	0	0.16	0
130_R2	Z3	73800650	2.02	0.03	0	0.27	0
130_R2	G9	66947236	6.97	0.1	0.3	1.04	0.01
130_R2	R9	85098918	9.54	0.14	0	1.12	0.02
130_R2	Z9	66225976	0.83	0.01	0	0.13	0
132	F0	82103488	97.8	203.96	96.24	11.91	24.84
132	G1	69999080	97.33	55.44	90.55	13.9	7.92
132	R1	59103578	95.16	15.89	95.75	16.1	2.69
132	Z1	58681304	97.85	234.05	96.45	16.68	39.88
132	G3	75210316	97.64	186.69	94.93	12.98	24.82
132	R3	72521542	96.89	50.3	96.18	13.36	6.94
132	Z3	63681826	97.94	483.38	96.59	15.38	75.91
132	G9	68089438	97.26	50.17	93.02	14.28	7.37
132	R9	74380490	96.01	32.61	95.48	12.91	4.38
132	Z9	70335360	97.88	276.78	96.35	13.92	39.35
135_R1	F0	83099226	98.85	52.62	97.56	11.9	6.33
135_R1	G1	67527456	98.5	43.92	85.13	14.59	6.5
135_R1	R1	57312146	94.84	12.14	97.24	16.55	2.12

Appendix 4 - Raw and normalised data input for statistical tests, MD-Campylobacter genomes used in Chapter 4

Stool_id	Conditions	reads_in	Breadth	Depth	Genome fraction	Breadth per 10M reads	Depth per 10M reads
135_R1	Z1	100468290	96.7	19.31	91.22	9.63	1.92
135_R1	G3	68813958	98.54	40.69	89.76	14.32	5.91
135_R1	R3	88619282	96.04	20.37	97.3	10.84	2.3
135_R1	Z3	61541608	98.01	14.75	95.77	15.93	2.4
135_R1	G9	76277592	98.5	43.56	91.82	12.91	5.71
135_R1	R9	87105596	96.85	18.65	97.25	11.12	2.14
135_R1	Z9	56988826	98.18	14.45	95.83	17.23	2.53
135_R2	F0	84836186	98.83	66.2	97.59	11.65	7.8
135_R2	G1	61460382	98.59	44.76	87.02	16.04	7.28
135_R2	R1	66398304	95.48	13.78	97.34	14.38	2.07
135_R2	Z1	56227236	98.34	16.55	97	17.49	2.94
135_R2	G3	68214230	98.67	48	90.96	14.46	7.04
135_R2	R3	74348210	95.48	13.78	96.72	12.84	1.85
135_R2	Z3	65725016	98.34	16.55	96.64	14.96	2.52
135_R2	G9	70058002	98.49	37.18	85.28	14.06	5.31
135_R2	R9	80963396	94.95	13.8	97.27	11.73	1.7
135_R2	Z9	65980016	98.29	20.81	96.57	14.9	3.15
136	F0	75702524	97.01	10.26	94.07	12.82	1.36
136	G1	69201452	92.39	4.89	23.11	13.35	0.71
136	R1	76040940	72.56	2.09	70.44	9.54	0.27
136	Z1	65752294	64.45	1.69	10.71	9.8	0.26
136	G3	60487756	89.86	4.05	27.97	14.86	0.67

Appendix 4 - Raw and normalised data input for statistical tests, MD-Campylobacter genomes used in Chapter 4

Stool_id	Conditions	reads_in	Breadth	Depth	Genome fraction	Breadth per 10M reads	Depth per 10M reads
136	R3	76671074	75.73	2.39	59.36	9.88	0.31
136	Z3	62930336	61.72	1.5	8.98	9.81	0.24
136	G9	58392242	90.2	3.97	36.3	15.45	0.68
136	R9	79650832	80.4	2.78	61.94	10.09	0.35
136	Z9	83596152	46.96	1.34	4.98	5.62	0.16
141	F0	72019188	98.28	12.32	94.77	13.65	1.71
141	G1	66464860	9.17	0.13	76.4	1.38	0.02
141	R1	60715036	25.39	0.4	0.03	4.18	0.07
141	Z1	122156238	94.04	6.69	0.19	7.7	0.55
141	G3	67956122	42.12	0.72	0	6.2	0.11
141	R3	62754532	4.83	0.07	1.12	0.77	0.01
141	Z3	84939984	58.3	1.65	6.2	6.86	0.19
141	G9	33139694	11.12	0.19	81.09	3.35	0.06
141	R9	69914624	94.86	7.19	0	13.57	1.03
141	Z9	51603250	67.48	2.13	10.21	13.08	0.41
143	F0	95747506	96.19	12.9	92.66	10.05	1.35
143	G1	71738480	97.64	77.38	96.22	13.61	10.79
143	R1	63123554	97.57	75.83	96.18	15.46	12.01
143	Z1	53263636	34.5	0.94	81.1	6.48	0.18
143	G3	54319056	97.52	72.34	96.32	17.95	13.32
143	R3	62222744	44	1.08	96.18	7.07	0.17
143	Z3	92744048	97.84	132.58	81.35	10.55	14.3

Appendix 4 - Raw and normalised data input for statistical tests, MD-Campylobacter genomes used in Chapter 4

Stool_id	Conditions	reads_in	Breadth	Depth	Genome fraction	Breadth per 10M reads	Depth per 10M reads
143	G9	65241864	97.37	55.97	96.21	14.92	8.58
143	R9	64561976	97.58	82.47	95.95	15.11	12.77
143	Z9	73541766	38.48	1.38	2.71	5.23	0.19
144	F0	60288958	44.04	0.85	2.54	7.3	0.14
144	G1	81772644	26.5	0.45	0.1	3.24	0.05
144	R1	56021382	20.97	0.32	0.59	3.74	0.06
144	Z1	68697528	14.84	0.23	0.06	2.16	0.03
144	G3	80289382	31.97	0.64	0.22	3.98	0.08
144	R3	73245362	21.4	0.4	0.4	2.92	0.06
144	Z3	67374118	12.17	0.23	0.06	1.81	0.03
144	G9	73004802	13.44	0.22	0	1.84	0.03
144	R9	59999222	1.41	0.02	0.01	0.24	0
144	Z9	70616660	12.86	0.24	0.06	1.82	0.03
145	F0	53999702	18.15	0.41	3.01	3.36	0.08
145	G1	72603804	3.43	0.05	0.03	0.47	0.01
145	R1	73871914	1.06	0.02	0	0.14	0
145	Z1	57553284	19.22	0.48	3.12	3.34	0.08
145	G3	68615464	3.2	0.06	0	0.47	0.01
145	R3	75194812	1.31	0.02	0	0.17	0
145	Z3	68065130	17.27	0.52	2.51	2.54	0.08
145	G9	78559188	1	0.02	0.04	0.13	0
145	R9	82092636	1.24	0.02	0	0.15	0

Appendix 4 - Raw and normalised data input for statistical tests, MD-Campylobacter genomes used in Chapter 4

Stool_id	Conditions	reads_in	Breadth	Depth	Genome fraction	Breadth per 10M reads	Depth per 10M reads
145	Z9	71914528	18.35	0.57	3.26	2.55	0.08
146	F0	68315170	89.65	3.96	61.79	13.12	0.58
146	G1	72603804	84.24	3.38	36.17	11.6	0.46
146	R1	73871914	75.8	2.67	49.18	10.26	0.36
146	Z1	60323412	44.22	0.97	1.37	7.33	0.16
146	G3	88507640	77.15	3.02	13.1	8.72	0.34
146	R3	78041878	62.58	1.83	34.29	8.02	0.23
146	Z3	71309860	44.9	1.03	1.73	6.3	0.14
146	G9	87628452	57.61	1.71	11.06	6.57	0.2
146	R9	69511646	56.67	1.56	11	8.15	0.22
146	Z9	81548292	51.31	1.29	3.61	6.29	0.16
147	F0	79012876	97.4	10.92	92.47	12.33	1.38
147	G1	77193912	79.39	3.16	84.28	10.28	0.41
147	R1	84882082	65.61	2.11	37.26	7.73	0.25
147	Z1	60089132	19.71	0.4	0.07	3.28	0.07
147	G3	73673330	86.14	4.13	36.02	11.69	0.56
147	R3	83262898	79.34	3.25	52.11	9.53	0.39
147	Z3	63019736	19.86	0.39	0.13	3.15	0.06
147	G9	83794960	54.12	1.41	4.49	6.46	0.17
147	R9	93174762	51.72	1.25	5.74	5.55	0.13
147	Z9	62647408	11.04	0.21	0	1.76	0.03

Appendix 5 - Full Shapiro-Wilk test for coverage metrics

Metric	Preservation Condition	Timepoint (No. of months)	N	W-statistic	p-value	Distribution
Breadth	G	1	13	0.854942441	0.033060536	non-normal
Breadth	G	3	13	0.923678219	0.281092525	normal
Breadth	G	9	13	0.866789281	0.047390513	non-normal
Breadth	R	1	13	0.916101158	0.222128898	normal
Breadth	R	3	13	0.883391559	0.079278894	normal
Breadth	R	9	13	0.901717544	0.141187534	normal
Breadth	Z	1	13	0.913752377	0.206361383	normal
Breadth	Z	3	13	0.910637319	0.187104166	normal
Breadth	Z	9	13	0.881864548	0.075582936	normal
Breadth	F	0	13	0.858258009	0.036541965	non-normal
Depth	G	1	13	0.69856751	0.00053692	non-normal
Depth	G	3	13	0.637623072	0.000141078	non-normal
Depth	G	9	13	0.700419903	0.000560301	non-normal
Depth	R	1	13	0.526682377	1.63E-05	non-normal
Depth	R	3	13	0.57060349	3.69E-05	non-normal
Depth	R	9	13	0.562690854	3.18E-05	non-normal
Depth	Z	1	13	0.362719357	1.08E-06	non-normal
Depth	Z	3	13	0.338255286	7.51E-07	non-normal
Depth	Z	9	13	0.374111712	1.29E-06	non-normal
Depth	F	0	13	0.563497603	3.22E-05	non-normal
Genome Fraction	G	1	13	0.802346349	0.007269113	non-normal
Genome Fraction	G	3	13	0.789987504	0.00519614	non-normal
Genome Fraction	G	9	13	0.811573148	0.009385944	non-normal
Genome Fraction	R	1	13	0.812190592	0.009549314	non-normal
Genome Fraction	R	3	13	0.862557054	0.041639403	non-normal
Genome Fraction	R	9	13	0.805896044	0.008016122	non-normal
Genome Fraction	Z	1	13	0.691936374	0.000461409	non-normal
Genome Fraction	Z	3	13	0.690853596	0.000450197	non-normal
Genome Fraction	Z	9	13	0.693484545	0.000477961	non-normal
Genome Fraction	F	0	13	0.737931728	0.00136644	non-normal

Appendix 5 - Full Shapiro-Wilk test for coverage metrics

Metric	Condition	Timepoint	N	W-statistic	p-value	Distribution
Log10 Breadth	G	1	13	0.840417683	0.021457007	non-normal
Log10 Breadth	G	3	13	0.786042094	0.00467538	non-normal
Log10 Breadth	G	9	13	0.830022931	0.015850354	non-normal
Log10 Breadth	R	1	13	0.781916022	0.004189848	non-normal
Log10 Breadth	R	3	13	0.809105575	0.008762167	non-normal
Log10 Breadth	R	9	13	0.796313405	0.00616456	non-normal
Log10 Breadth	Z	1	13	0.797406554	0.006350596	non-normal
Log10 Breadth	Z	3	13	0.832462668	0.017009644	non-normal
Log10 Breadth	Z	9	13	0.820751965	0.012154821	non-normal
Log10 Breadth	F	0	13	0.734928131	0.001269672	non-normal
Log10 Depth	G	1	13	0.906135619	0.162337244	normal
Log10 Depth	G	3	13	0.950947285	0.61271143	normal
Log10 Depth	G	9	13	0.931608677	0.35767749	normal
Log10 Depth	R	1	13	0.967500985	0.862776041	normal
Log10 Depth	R	3	13	0.962086797	0.785533309	normal
Log10 Depth	R	9	13	0.955403507	0.681881964	normal
Log10 Depth	Z	1	13	0.94740057	0.559474766	normal
Log10 Depth	Z	3	13	0.951266468	0.617599607	normal
Log10 Depth	Z	9	13	0.963506341	0.806693971	normal
Log10 Depth	F	0	13	0.938071132	0.432416171	normal
Log10 Genome Fraction	G	1	11	0.708308935	0.000598396	non-normal
Log10 Genome Fraction	G	3	9	0.77873224	0.011646295	non-normal
Log10 Genome Fraction	G	9	12	0.833698273	0.023229852	non-normal
Log10 Genome Fraction	R	1	11	0.72138226	0.000882791	non-normal
Log10 Genome Fraction	R	3	10	0.685850978	0.000593836	non-normal
Log10 Genome Fraction	R	9	9	0.672017217	0.00066825	non-normal
Log10 Genome Fraction	Z	1	11	0.870460033	0.078584388	normal
Log10 Genome Fraction	Z	3	11	0.875127494	0.090187795	normal
Log10 Genome Fraction	Z	9	10	0.878285885	0.124703094	normal
Log10 Genome Fraction	F	0	11	0.566970944	9.92E-06	non-normal

Appendix 6 - MD-*Campylobacter* genomes GTDB-Tk classification, and mean qPCR results used in Chapter 4

MDG	ST Score	Complete ST	*Campylobacter_D jejuni*	*CadF* mean Cp	Human mean Cp
124_TP0	**7**	**yes**	yes	25.23	33.39
124_R_1M	**0**	**no**	*Campylobacter_D coli*	29.53	33.42
124_R_3M	**1**	**no**	*Campylobacter_D hepaticus*	30.82	34.03
124_R_9M	**1**	**no**	*Campylobacter_D;s__*	31.2	36.47
124_G_1M	**3**	**no**	*Campylobacter_D;s__*	26.59	32.65
124_G_3M	**7**	**yes**	yes	29.1	34.67
124_G_9M	**6**	**no**	yes	29.37	34.07
124_Z_1M	**1**	**no**	*Campylobacter_D coli*	26.97	32.69
124_Z_3M	**6**	**no**	yes	28.96	33.77
124_Z_9M	**1**	**no**	yes	26.94	32.58
130r1_TP0	**0**	**no**	Unclassified	34.38	29.87
130r1_R_1M	**0**	**no**	Unclassified	35.32	30.21
130r1_R_3M	**0**	**no**	Unclassified	37.13	29.83
130r1_R_9M	**0**	**no**	Unclassified	36.74	29.99
130r1_G_1M	**0**	**no**	Unclassified	36.79	30.64
130r1_G_3M	**0**	**no**	Unclassified	35.93	29.86
130r1_G_9M	**0**	**no**	Unclassified	37.27	30.11
130r1_Z_1M	**0**	**no**	Unclassified	35.71	25.85
130r1_Z_3M	**0**	**no**	Unclassified	35.75	25.39
130r1_Z_9M	**0**	**no**	Unclassified Bacteria	36.86	25.64
130r2_TP0	**0**	**no**	Unclassified	34.37	29.1
130r2_R_1M	**0**	**no**	Unclassified Bacteria	35.46	29.87
130r2_R_3M	**0**	**no**	Unclassified Bacteria	37.09	30.27
130r2_R_9M	**0**	**no**	Unclassified Bacteria	36.77	29.78
130r2_G_1M	**0**	**no**	Unclassified Bacteria	36.83	30.78
130r2_G_3M	**0**	**no**	Unclassified Bacteria	35.91	29.81
130r2_G_9M	**0**	**no**	Unclassified Bacteria	37	30.21
130r2_Z_1M	**0**	**no**	Unclassified	35.73	25.72
130r2_Z_3M	**0**	**no**	Unclassified	35.75	25.48
130r2_Z_9M	**0**	**no**	Unclassified	36.82	25.67
132_TP0	**7**	**yes**	yes	23.21	32.67
132_R_1M	**7**	**yes**	yes	27.58	33.63
132_R_3M	**7**	**yes**	yes	27.48	33.64
132_R_9M	**7**	**yes**	yes	22.99	31.88
132_G_1M	**7**	**yes**	yes	26.27	32.98
132_G_3M	**7**	**yes**	yes	25.39	33.03
132_G_9M	**7**	**yes**	yes	27.01	32.83
132_Z_1M	**7**	**yes**	yes	21.59	30.23

Appendix 6 - MD-*Campylobacter* genomes GTDB-Tk classification, and mean qPCR results used in Chapter 4

MDG	ST Score	Complete ST	*Campylobacter_D jejuni*	*CadF* mean Cp	Human mean Cp
132_Z_3M	**7**	**yes**	yes	19.85	27.7
132_Z_9M	**7**	**yes**	yes	19.15	28.01
135r1_TP0	**7**	**yes**	yes	23	33.29
135r1_R_1M	**7**	**yes**	yes	26.11	32.21
135r1_R_3M	**7**	**yes**	yes	25.64	32.8
135r1_R_9M	**7**	**yes**	yes	25.92	33.72
135r1_G_1M	**7**	**yes**	yes	24.69	33.52
135r1_G_3M	**7**	**yes**	yes	25.14	34.11
135r1_G_9M	**7**	**yes**	yes	25.34	34.25
135r1_Z_1M	**7**	**yes**	yes	25.05	31.57
135r1_Z_3M	**7**	**yes**	yes	23.79	30.71
135r1_Z_9M	**7**	**yes**	yes	24.04	30.84
135r2_TP0	**7**	**yes**	yes	22.76	33.84
135r2_R_1M	**7**	**yes**	yes	25.96	32.44
135r2_R_3M	**7**	**yes**	yes	25.31	32.13
135r2_R_9M	**7**	**yes**	yes	26.87	34.17
135r2_G_1M	**7**	**yes**	yes	24.71	33.6
135r2_G_3M	**7**	**yes**	yes	24.24	33.85
135r2_G_9M	**7**	**yes**	yes	24.12	34.04
135r2_Z_1M	**7**	**yes**	yes	24.54	31.3
135r2_Z_3M	**7**	**yes**	yes	24.59	31.25
135r2_Z_9M	**7**	**yes**	yes	24.75	31.25
136_TP0	**6**	**no**	yes	25.5	30.68
136_R_1M	**3**	**no**	*Campylobacter_D jejuni_D*	29.75	30.5
136_R_3M	**0**	**no**	*Campylobacter_D jejuni_D*	30.1	30.97
136_R_9M	**4**	**no**	*Campylobacter_D jejuni_D*	30.78	32.11
136_G_1M	**3**	**no**	yes	28.14	31.2
136_G_3M	**4**	**no**	yes	29.28	31.26
136_G_9M	**4**	**no**	yes	28.75	31.58
136_Z_1M	**0**	**no**	*Campylobacter_D jejuni_D*	26.77	31.42
136_Z_3M	**0**	**no**	*Campylobacter_D coli*	28.01	31.11
136_Z_9M	**0**	**no**	Unclassified Bacteria	26.65	30.9
141_TP0	**7**	**yes**	yes	26.1	26.4
141_R_1M	**6**	**no**	yes	29.62	26.72
141_R_3M	**0**	**no**	Unclassified	28.65	22.69
141_R_9M	**7**	**no**	yes	27.17	27.31
141_G_1M	**0**	**no**	Unclassified Bacteria	30.22	23.51

Appendix 6 - MD-*Campylobacter* genomes GTDB-Tk classification, and mean qPCR results used in Chapter 4

MDG	ST Score	Complete ST	*Campylobacter_D jejuni*	*CadF* mean Cp	Human mean Cp
141_G_3M	**1**	**no**	Unclassified Bacteria	28.52	24.15
141_G_9M	**0**	**no**	Unclassified Bacteria	35.77	28.01
141_Z_1M	**0**	**no**	Unclassified	29.02	23.9
141_Z_3M	**0**	**no**	Unclassified Bacteria	28.02	23.95
141_Z_9M	**0**	**no**	*Campylobacter_D;s__*	25.84	25.51
143_TP0	**7**	**yes**	yes	25.98	24.75
143_R_1M	**7**	**yes**	yes	25.45	31
143_R_3M	**7**	**yes**	yes	24.3	30.65
143_R_9M	**7**	**yes**	yes	24.09	30.97
143_G_1M	**7**	**yes**	yes	26.19	31.6
143_G_3M	**7**	**yes**	yes	25.25	31.28
143_G_9M	**7**	**yes**	yes	26.43	31.55
143_Z_1M	**0**	**no**	Unclassified Archaea	26.81	23.59
143_Z_3M	**0**	**no**	Unclassified Bacteria	36.7	23.69
143_Z_9M	**0**	**no**	Unclassified Archaea	33.44	27.98
144_TP0	**0**	**no**	Unclassified Bacteria	32.64	32.83
144_R_1M	**0**	**no**	Unclassified Bacteria	34.23	32.62
144_R_3M	**0**	**no**	Unclassified	31.17	31.14
144_R_9M	**0**	**no**	Unclassified	39.21	32.42
144_G_1M	**0**	**no**	Unclassified Bacteria	33.45	31.99
144_G_3M	**0**	**no**	Unclassified	0	32.06
144_G_9M	**0**	**no**	Unclassified Bacteria	35.72	31.87
144_Z_1M	**0**	**no**	Unclassified Archaea	31.34	28.36
144_Z_3M	**0**	**no**	Unclassified	30.61	28.78
144_Z_9M	**0**	**no**	Unclassified	31.51	29.55
145_TP0	**0**	**no**	Unclassified Bacteria	0	32.97
145_R_1M	**0**	**no**	Unclassified	40	32.64
145_R_3M	**0**	**no**	Unclassified	36.86	31.39
145_R_9M	**0**	**no**	Unclassified	36.66	30.55
145_G_1M	**0**	**no**	Unclassified	35.12	32.76
145_G_3M	**0**	**no**	Unclassified	33.49	31.74
145_G_9M	**0**	**no**	Unclassified	35.74	31.01
145_Z_1M	**0**	**no**	Unclassified Bacteria	31.93	32.56
145_Z_3M	**0**	**no**	Unclassified Bacteria	28.33	30.56
145_Z_9M	**0**	**no**	Unclassified Bacteria	0	30.72
146_TP0	**6**	**no**	yes	29.59	28.01
146_R_1M	**2**	**no**	yes	31.94	30.68
146_R_3M	**0**	**no**	*Campylobacter_D;s__*	30	27.65
146_R_9M	**1**	**no**	*Campylobacter_D;s__*	30.77	28.06

Appendix 6 - MD-*Campylobacter* genomes GTDB-Tk classification, and mean qPCR results used in Chapter 4

MDG	ST Score	Complete ST	*Campylobacter_D jejuni*	*CadF* mean Cp	Human mean Cp
146_G_1M	**4**	**no**	yes	31.15	31.62
146_G_3M	**3**	**no**	yes	32.49	29.97
146_G_9M	**2**	**no**	*Campylobacter_D;s__*	31.26	29.64
146_Z_1M	**0**	**no**	Unclassified Bacteria	32.04	24.87
146_Z_3M	**0**	**no**	Unclassified Bacteria	32.73	25.86
146_Z_9M	**0**	**no**	Unclassified Bacteria	30.73	25.87
147_TP0	**7**	**yes**	yes	27.01	29.68
147_R_1M	**3**	**no**	*Campylobacter_D;s__*	31.19	29.37
147_R_3M	**4**	**no**	yes	32.45	30.21
147_R_9M	**0**	**no**	Unclassified Bacteria	31.55	27.84
147_G_1M	**2**	**no**	yes	30.86	29.89
147_G_3M	**6**	**no**	yes	28.21	28.64
147_G_9M	**0**	**no**	Unclassified Bacteria	31.83	27.91
147_Z_1M	**0**	**no**	Unclassified Bacteria	27.54	24.52
147_Z_3M	**0**	**no**	Unclassified	27.11	24.75
147_Z_9M	**0**	**no**	Unclassified	28.8	24.68
144_R_9M	**0**	**no**	Unclassified	39.21	32.42
144_G_1M	**0**	**no**	Unclassified Bacteria	33.45	31.99
144_G_3M	**0**	**no**	Unclassified	0	32.06
144_G_9M	**0**	**no**	Unclassified Bacteria	35.72	31.87
144_Z_1M	**0**	**no**	Unclassified Archaea	31.34	28.36
144_Z_3M	**0**	**no**	Unclassified	30.61	28.78
144_Z_9M	**0**	**no**	Unclassified	31.51	29.55
145_TP0	**0**	**no**	Unclassified Bacteria	0	32.97
145_R_1M	**0**	**no**	Unclassified	40	32.64
145_R_3M	**0**	**no**	Unclassified	36.86	31.39
145_R_9M	**0**	**no**	Unclassified	36.66	30.55
145_G_1M	**0**	**no**	Unclassified	35.12	32.76
145_G_3M	**0**	**no**	Unclassified	33.49	31.74
145_G_9M	**0**	**no**	Unclassified	35.74	31.01
145_Z_1M	**0**	**no**	Unclassified Bacteria	31.93	32.56
145_Z_3M	**0**	**no**	Unclassified Bacteria	28.33	30.56
145_Z_9M	**0**	**no**	Unclassified Bacteria	0	30.72
146_TP0	**6**	**no**	yes	29.59	28.01
146_R_1M	**2**	**no**	yes	31.94	30.68
146_R_3M	**0**	**no**	*Campylobacter_D;s__*	30	27.65
146_R_9M	**1**	**no**	*Campylobacter_D;s__*	30.77	28.06
146_G_1M	**4**	**no**	yes	31.15	31.62
146_G_3M	**3**	**no**	yes	32.49	29.97

Appendix 6 - MD-*Campylobacter* genomes GTDB-Tk classification, and mean qPCR results used in Chapter 4

MDG	ST Score	Complete ST	*Campylobacter_D jejuni*	*CadF* mean Cp	Human mean Cp
146_G_9M	**2**	**no**	*Campylobacter_D;s__*	31.26	29.64
146_Z_1M	**0**	**no**	Unclassified Bacteria	32.04	24.87
146_Z_3M	**0**	**no**	Unclassified Bacteria	32.73	25.86
146_Z_9M	**0**	**no**	Unclassified Bacteria	30.73	25.87
147_TP0	**7**	**yes**	yes	27.01	29.68
147_R_1M	**3**	**no**	*Campylobacter_D;s__*	31.19	29.37
147_R_3M	**4**	**no**	yes	32.45	30.21
147_R_9M	**0**	**no**	Unclassified Bacteria	31.55	27.84
147_G_1M	**2**	**no**	yes	30.86	29.89
147_G_3M	**6**	**no**	yes	28.21	28.64
147_G_9M	**0**	**no**	Unclassified Bacteria	31.83	27.91
147_Z_1M	**0**	**no**	Unclassified Bacteria	27.54	24.52
147_Z_3M	**0**	**no**	Unclassified	27.11	24.75
147_Z_9M	**0**	**no**	Unclassified	28.8	24.68
147_Z_9M	**0**	**no**	Unclassified	28.8	24.68

Appendix 7 - Isolate and MD-*Campylobacter* genomes full MLST scores used in Chapter 4

Isolate	MLST scheme	ST	aspA	glnA	gltA	glyA	pgm	tkt	uncA
124-6	campylobacter	464	24	2	2	2	10	3	1
124-1	campylobacter	464	24	2	2	2	10	3	1
124-7	campylobacter	464	24	2	2	2	10	3	1
124-10	campylobacter	464	24	2	2	2	10	3	1
124-8	campylobacter	464	24	2	2	2	10	3	1
124-3	campylobacter	464	24	2	2	2	10	3	1
124-11	campylobacter	464	24	2	2	2	10	3	1
124-5	campylobacter	464	24	2	2	2	10	3	1
124-12	campylobacter	464	24	2	2	2	10	3	1
124-2	campylobacter	464	24	2	2	2	10	3	1
MDG	MLST scheme	ST	aspA	glnA	gltA	glyA	pgm	tkt	uncA
124_G_1M	campylobacter	-	~24	683?	594?	2	956?	~3	-
124_G_3M	campylobacter	464	24	2	2	2	10	3	1
124_G_9M	campylobacter	-	24	683?	2	2	10	3	1?
124_R_1M	campylobacter	-	-	-	594?	-	956?	-	-
124_R_3M	campylobacter	-	504?	671?	594?	767?	956?	-	1
124_R_9M	campylobacter	-	504?	694?	2	-	956?	711?	615?
124_TP0	campylobacter	464	24	2	2	2	10	3	1
124_Z_1M	campylobacter	-	509?	-	~2	767?	518?	-	615?
124_Z_3M	campylobacter	-	24	~2	~2	372?	10	3	1
124_Z_9M	campylobacter	-	509?	695?	594?	108?	10?	741?	615?

Appendix 7 - Isolate and MD-*Campylobacter* genomes full MLST scores used in Chapter 4

Isolate	MLST scheme	ST	aspA	glnA	gltA	glyA	pgm	tkt	uncA
130-2	campylobacter	-	7	97	594?	2	135	68	26
130-6	campylobacter	791	7	97	5	2	135	68	26
130-3	campylobacter	791	7	97	5	2	135	68	26
130-8	campylobacter	791	7	97	5	2	135	68	26
130-4	campylobacter	791	7	97	5	2	135	68	26
130-10	campylobacter	791	7	97	5	2	135	68	26
130-7	campylobacter	791	7	97	5	2	135	68	26
130-1	campylobacter	791	7	97	5	2	135	68	26
130-9	campylobacter	791	7	97	5	2	135	68	26
MDG	MLST scheme	ST	aspA	glnA	gltA	glyA	pgm	tkt	uncA
130r1_Z_9M	campylobacter	-	-	-	-	-	-	-	-
130r1_Z_1M	campylobacter	-	-	-	-	-	-	-	-
130r1_Z_3M	campylobacter	-	-	-	-	-	-	-	-
130r1_R_1M	campylobacter	-	-	-	-	-	-	-	-
130r1_G_1M	campylobacter	-	-	-	-	-	-	-	-
130r1_R_3M	campylobacter	-	-	-	-	-	-	-	-
130r1_G_9M	campylobacter	-	-	-	-	-	-	-	-
130r1_R_9M	campylobacter	-	-	-	-	-	-	-	-
130r1_TP0	campylobacter	-	-	-	-	-	-	-	-
130r1_G_3M	campylobacter	-	-	-	-	-	-	-	-
130r2_Z_9M	campylobacter	-	-	-	-	-	-	-	-
130r2_Z_1M	campylobacter	-	-	-	-	-	-	-	-
130r2_Z_3M	campylobacter	-	-	-	-	-	-	-	-
130r2_R_3M	campylobacter	-	-	-	-	-	-	-	-
130r2_G_9M	campylobacter	-	-	-	-	-	-	-	-
130r2_G_3M	campylobacter	-	-	-	578?	-	-	-	-
130r2_TP0	campylobacter	-	-	-	-	-	-	-	-
130r2_R_9M	campylobacter	-	-	-	-	-	-	-	-
130r2_R_1M	campylobacter	-	-	-	-	-	-	-	-
130r2_G_1M	campylobacter	-	-	-	-	-	-	-	-

Appendix 7 - Isolate and MD-*Campylobacter* genomes full MLST scores used in Chapter 4

Isolate	MLST scheme	ST	aspA	glnA	gltA	glyA	pgm	tkt	uncA
132C-8	campylobacter	-	2	17	5	2	10	12	6
132C-11	campylobacter	-	2	17	5	2	10	12	6
132C-3	campylobacter	-	2	17	5	2	10	12	6
132C-1	campylobacter	-	2	17	5	2	10	12	6
132C-7	campylobacter	-	2	17	5	2	10	12	6
132C-9	campylobacter	-	2	17	5	2	10	12	6
132C-6	campylobacter	-	2	17	5	2	10	12	6
132C-10	campylobacter	-	2	17	5	2	10	12	6
132C-2	campylobacter	-	2	17	5	2	10	12	6
132C-4	campylobacter	-	2	17	5	2	10	12	6
132C-5	campylobacter	-	2	17	5	2	10	12	6
132C-12	campylobacter	-	2	17	5	2	10	12	6
MDG	MLST scheme	ST	aspA	glnA	gltA	glyA	pgm	tkt	uncA
132_R_1M	campylobacter	-	2	17	5	2	10	12	6
132_R_9M	campylobacter	-	2	17	5	2	10	12	6
132_G_9M	campylobacter	-	2	17	5	2	10	12	6
132_R_3M	campylobacter	-	2	17	5	2	10	12	6
132_G_1M	campylobacter	-	2	17	5	2	10	12	6
132_G_3M	campylobacter	-	2	17	5	2	10	12	6
132_Z_1M	campylobacter	-	2	17	5	2	10	12	6
132_TP0	campylobacter	-	2	17	5	2	10	12	6
132_Z_9M	campylobacter	-	2	17	5	2	10	12	6
132_Z_3M	campylobacter	-	2	17	5	2	10	12	6

Appendix 7 - Isolate and MD-*Campylobacter* genomes full MLST scores used in Chapter 4

Isolate	MLST scheme	ST	aspA	glnA	gltA	glyA	pgm	tkt	uncA
135C-6	campylobacter	1707	9	2	5	2	11	3	1
135C-7	campylobacter	1707	9	2	5	2	11	3	1
135C-1	campylobacter	1707	9	2	5	2	11	3	1
135C-3	campylobacter	1707	9	2	5	2	11	3	1
135C-9	campylobacter	1707	9	2	5	2	11	3	1
135C-2	campylobacter	1707	9	2	5	2	11	3	1
135C-8	campylobacter	1707	9	2	5	2	11	3	1
135C-4	campylobacter	1707	9	2	5	2	11	3	1
135C-10	campylobacter	1707	9	2	5	2	11	3	1
135C-11	campylobacter	1707	9	2	5	2	11	3	1
135C-5	campylobacter	1707	9	2	5	2	11	3	1
135C-12	campylobacter	1707	9	2	5	2	11	3	1
MDG	MLST scheme	ST	aspA	glnA	gltA	glyA	pgm	tkt	uncA
135r1_R_1M	campylobacter	1707	9	2	5	2	11	3	1
135r1_G_1M	campylobacter	1707	9	2	5	2	11	3	1
135r1_Z_3M	campylobacter	1707	9	2	5	2	11	3	1
135r1_Z_9M	campylobacter	1707	9	2	5	2	11	3	1
135r1_Z_1M	campylobacter	1707	9	2	5	2	11	3	1
135r1_G_9M	campylobacter	1707	9	2	5	2	11	3	1
135r1_R_3M	campylobacter	1707	9	2	5	2	11	3	1
135r1_R_9M	campylobacter	1707	9	2	5	2	11	3	1
135r1_G_3M	campylobacter	1707	9	2	5	2	11	3	1
135r1_TP0	campylobacter	1707	9	2	5	2	11	3	1
135r2_Z_1M	campylobacter	1707	9	2	5	2	11	3	1
135r2_R_9M	campylobacter	1707	9	2	5	2	11	3	1
135r2_G_3M	campylobacter	1707	9	2	5	2	11	3	1
135r2_G_9M	campylobacter	1707	9	2	5	2	11	3	1
135r2_R_3M	campylobacter	1707	9	2	5	2	11	3	1
135r2_G_1M	campylobacter	1707	9	2	5	2	11	3	1
135r2_R_1M	campylobacter	1707	9	2	5	2	11	3	1
135r2_Z_3M	campylobacter	1707	9	2	5	2	11	3	1
135r2_Z_9M	campylobacter	1707	9	2	5	2	11	3	1
135r2_TP0	campylobacter	1707	9	2	5	2	11	3	1

Appendix 7 - Isolate and MD-*Campylobacter* genomes full MLST scores used in Chapter 4

Isolate	MLST scheme	ST	aspA	glnA	gltA	glyA	pgm	tkt	uncA
136C-6	campylobacter	4697	8	17	5	2	10	3	6
136C-8	campylobacter	4697	8	17	5	2	10	3	6
136C-12	campylobacter	4697	8	17	5	2	10	3	6
136C-4	campylobacter	4697	8	17	5	2	10	3	6
136C-2	campylobacter	4697	8	17	5	2	10	3	6
136C-10	campylobacter	4697	8	17	5	2	10	3	6
MDG	MLST scheme	ST	aspA	glnA	gltA	glyA	pgm	tkt	uncA
136_Z_3M	campylobacter	-	-	-	-	-	-	-	-
136_Z_9M	campylobacter	-	-	-	-	-	-	-	615?
136_G_1M	campylobacter	-	8	694?	5	747?	956?	386?	6
136_R_1M	campylobacter	-	495?	~17	5	775?	943?	~3	615?
136_R_9M	campylobacter	-	504?	~17	5	747?	948?	3	6
136_G_3M	campylobacter	-	8	17	578?	775?	956?	3	6
136_G_9M	campylobacter	-	510?	17	578?	2	956?	3	6
136_R_3M	campylobacter	-	504?	-	-	747?	956?	397?	614?
136_Z_1M	campylobacter	-	-	-	-	775?	948?	-	-
136_TP0	campylobacter	-	8	17	5	~2	10	733?	6

Appendix 7 - Isolate and MD-*Campylobacter* genomes full MLST scores used in Chapter 4

Isolate	MLST scheme	ST	aspA	glnA	gltA	glyA	pgm	tkt	uncA
141C-10	campylobacter	-	2	21	12	62	11	67	6
141C-1	campylobacter	-	2	21	12	62	11	67	6
141C-12	campylobacter	-	2	21	12	62	11	67	6
141C-3	campylobacter	-	2	21	12	62	11	67	6
141C-6	campylobacter	-	2	21	12	62	11	67	6
141C-8	campylobacter	-	2	21	12	62	11	67	6
141C-7	campylobacter	-	2	21	12	62	11	67	6
141C-4	campylobacter	-	2	21	12	62	11	67	6
141C-9	campylobacter	-	2	21	12	62	11	67	6
141C-11	campylobacter	-	2	21	12	62	11	67	6
141C-2	campylobacter	-	2	21	12	62	11	67	6
141C-5	campylobacter	-	2	21	12	62	11	67	6
MDG	MLST scheme	ST	aspA	glnA	gltA	glyA	pgm	tkt	uncA
141_R_3M	campylobacter	-	-	-	-	-	-	-	-
141_G_1M	campylobacter	-	-	-	-	-	-	-	-
141_G_9M	campylobacter	-	-	-	-	-	-	-	-
141_Z_1M	campylobacter	-	-	-	-	-	-	-	-
141_G_3M	campylobacter	-	-	669?	12	-	-	-	615?
141_Z_3M	campylobacter	-	-	-	591?	-	-	-	-
141_Z_9M	campylobacter	-	510?	694?	-	-	-	-	-
141_R_1M	campylobacter	-	~2	21	420?	62	11	67	6
141_R_9M	campylobacter	-	2	21	12	62	~11	67?	6
141_TP0	campylobacter	-	2	21	12	62	11	67	6

Appendix 7 - Isolate and MD-*Campylobacter* genomes full MLST scores used in Chapter 4

Isolate	MLST scheme	ST	aspA	glnA	gltA	glyA	pgm	tkt	uncA
143C-8	campylobacter	21	2	1	1	3	2	1	5
143C-3	campylobacter	21	2	1	1	3	2	1	5
143C-10	campylobacter	21	2	1	1	3	2	1	5
143C-4	campylobacter	21	2	1	1	3	2	1	5
143C-6	campylobacter	21	2	1	1	3	2	1	5
143C-5	campylobacter	21	2	1	1	3	2	1	5
143C-9	campylobacter	21	2	1	1	3	2	1	5
143C-11	campylobacter	21	2	1	1	3	2	1	5
143C-7	campylobacter	21	2	1	1	3	2	1	5
143C-12	campylobacter	21	2	1	1	3	2	1	5
143C-1	campylobacter	21	2	1	1	3	2	1	5
143C-2	campylobacter	21	2	1	1	3	2	1	5
MDG	MLST scheme	ST	aspA	glnA	gltA	glyA	pgm	tkt	uncA
143_R_1M	campylobacter	21	2	1	1	3	2	1	5
143_G_1M	campylobacter	21	2	1	1	3	2	1	5
143_Z_3M	campylobacter	-	510?	-	591?	-	-	-	-
143_R_9M	campylobacter	21	2	1	1	3	2	1	5
143_G_9M	campylobacter	21	2	1	1	3	2	1	5
143_Z_9M	campylobacter	-	-	-	-	-	-	-	596?
143_TP0	campylobacter	21	2	1	1	3	2	1	5
143_Z_1M	campylobacter	-	-	-	-	-	-	-	614?
143_R_3M	campylobacter	21	2	1	1	3	2	1	5
143_G_3M	campylobacter	21	2	1	1	3	2	1	5

Appendix 7 - Isolate and MD-*Campylobacter* genomes full MLST scores used in Chapter 4

Isolate	MLST scheme	ST	aspA	glnA	gltA	glyA	pgm	tkt	uncA
144C-12	campylobacter	6175	2	1	5	10	608	1	5
144C-8	campylobacter	6175	2	1	5	10	608	1	5
144C-6	campylobacter	6175	2	1	5	10	608	1	5
144C-3	campylobacter	6175	2	1	5	10	608	1	5
144C-1	campylobacter	6175	2	1	5	10	608	1	5
144C-7	campylobacter	6175	2	1	5	10	608	1	5
144C-11	campylobacter	6175	2	1	5	10	608	1	5
144C-10	campylobacter	6175	2	1	5	10	608	1	5
144C-4	campylobacter	6175	2	1	5	10	608	1	5
144C-2	campylobacter	6175	2	1	5	10	608	1	5
144C-5	campylobacter	6175	2	1	5	10	608	1	5
144C-9	campylobacter	6175	2	1	5	10	608	1	5
MDG	MLST scheme	ST	aspA	glnA	gltA	glyA	pgm	tkt	uncA
144_Z_1M	campylobacter	-	-	-	-	-	-	-	-
144_TP0	campylobacter	-	-	616?	594?	-	-	-	523?
144_G_3M	campylobacter	-	-	-	559?	-	-	-	-
144_R_3M	campylobacter	-	-	-	-	-	-	-	-
144_Z_3M	campylobacter	-	-	-	-	-	-	-	-
144_G_9M	campylobacter	-	-	-	-	-	-	-	596?
144_R_9M	campylobacter	-	-	-	-	-	-	-	-
144_R_1M	campylobacter	-	-	-	-	-	-	-	-
144_Z_9M	campylobacter	-	-	-	-	-	-	-	615?
144_G_1M	campylobacter	-	-	-	-	-	-	741?	523?

Appendix 7 - Isolate and MD-*Campylobacter* genomes full MLST scores used in Chapter 4

Isolate	MLST scheme	ST	aspA	glnA	gltA	glyA	pgm	tkt	uncA
145C-7	campylobacter	829	33	39	30	82	113	43	17
145C-4	campylobacter	829	33	39	30	82	113	43	17
145C-11	campylobacter	829	33	39	30	82	113	43	17
145C-12	campylobacter	829	33	39	30	82	113	43	17
145C-8	campylobacter	829	33	39	30	82	113	43	17
145C-5	campylobacter	829	33	39	30	82	113	43	17
145C-6	campylobacter	829	33	39	30	82	113	43	17
145C-3	campylobacter	829	33	39	30	82	113	43	17
145C-1	campylobacter	829	33	39	30	82	113	43	17
145C-10	campylobacter	829	33	39	30	82	113	43	17
145C-2	campylobacter	829	33	39	30	82	113	43	17
MDG	MLST scheme	ST	aspA	glnA	gltA	glyA	pgm	tkt	uncA
145_G_1M	campylobacter	-	-	-	-	-	-	-	-
145_R_1M	campylobacter	-	-	-	-	-	-	-	-
145_Z_3M	campylobacter	-	-	-	541?	-	-	-	-
145_TP0	campylobacter	-	-	-	-	-	-	-	-
145_R_3M	campylobacter	-	-	-	-	-	-	-	-
145_G_3M	campylobacter	-	482?	-	-	-	-	-	-
145_Z_9M	campylobacter	-	-	-	541?	-	-	-	-
145_R_9M	campylobacter	-	-	-	-	-	-	-	-
145_G_9M	campylobacter	-	-	-	-	-	-	-	-
145_Z_1M	campylobacter	-	508?	-	-	-	-	-	-

Appendix 7 - Isolate and MD-*Campylobacter* genomes full MLST scores used in Chapter 4

Isolate	MLST scheme	ST	aspA	glnA	gltA	glyA	pgm	tkt	uncA
146C-12	campylobacter	19	2	1	5	3	2	1	5
146C-8	campylobacter	19	2	1	5	3	2	1	5
146C-1	campylobacter	19	2	1	5	3	2	1	5
146C-5	campylobacter	19	2	1	5	3	2	1	5
146C-2	campylobacter	19	2	1	5	3	2	1	5
146C-9	campylobacter	19	2	1	5	3	2	1	5
146C-10	campylobacter	19	2	1	5	3	2	1	5
146C-11	campylobacter	19	2	1	5	3	2	1	5
146C-6	campylobacter	19	2	1	5	3	2	1	5
MDG	MLST scheme	ST	aspA	glnA	gltA	glyA	pgm	tkt	uncA
146_R_9M	campylobacter	-	510?	392?	5	772?	-	711?	523?
146_G_9M	campylobacter	-	510?	-	578?	772?	898?	1	5
146_Z_9M	campylobacter	-	-	-	-	-	-	-	-
146_R_3M	campylobacter	-	489?	616?	578?	-	-	-	523?
146_G_3M	campylobacter	-	43?	1	5	714?	898?	711?	5
146_Z_3M	campylobacter	-	-	-	-	-	-	732?	-
146_TP0	campylobacter	-	2	~1	~5	765?	2	~1	5
146_Z_1M	campylobacter	-	-	-	-	-	-	-	523?
146_R_1M	campylobacter	-	308?	694?	578?	767?	2	1	523?
146_G_1M	campylobacter	-	-	1	5	772?	957?	1	5

Appendix 7 - Isolate and MD-*Campylobacter* genomes full MLST scores used in Chapter 4

Isolate	MLST scheme	ST	aspA	glnA	gltA	glyA	pgm	tkt	uncA
147C-9	campylobacter	400	8	17	5	2	10	59	6
147C-1	campylobacter	400	8	17	5	2	10	59	6
147C-5	campylobacter	400	8	17	5	2	10	59	6
147C-2	campylobacter	400	8	17	5	2	10	59	6
147C-7	campylobacter	400	8	17	5	2	10	59	6
147C-10	campylobacter	400	8	17	5	2	10	59	6
147C-8	campylobacter	400	8	17	5	2	10	59	6
147C-11	campylobacter	400	8	17	5	2	10	59	6
147C-3	campylobacter	400	8	17	5	2	10	59	6
147C-12	campylobacter	400	8	17	5	2	10	59	6
147C-6	campylobacter	400	8	17	5	2	10	59	6
147C-4	campylobacter	400	8	17	5	2	10	59	6
MDG	MLST scheme	ST	aspA	glnA	gltA	glyA	pgm	tkt	uncA
147_Z_9M	campylobacter	-	-	-	578?	-	-	-	-
147_G_9M	campylobacter	-	-	-	-	-	956?	-	585?
147_R_9M	campylobacter	-	504?	-	-	-	948?	-	-
147_Z_3M	campylobacter	-	-	-	-	-	-	-	-
147_G_3M	campylobacter	-	8	17	~5	2	10	59	6
147_R_3M	campylobacter	-	8	-	5	2	167?	137?	6
147_TP0	campylobacter	400	8	17	5	2	10	59	6
147_G_1M	campylobacter	-	439?	17	578?	-	956?	59	585?
147_R_1M	campylobacter	-	-	551?	5	2	956?	~59	6
147_Z_1M	campylobacter	-	-	-	-	-	-	-	-

Appendix 8 - CheckM results for MD-*Campylobacter* genomes, these values are not standardised to reads in and represent the full sequencing yield of each sample

Stool Id	Condition	Timepoint	Completeness	Contamination	Strain heterogeneity
124	F	0	88.59	2.39	6.25
124	G	1	33.84	1.45	0
124	G	3	91.7	3.83	20.83
124	G	9	81.34	4.5	16
124	R	1	13.18	0.21	66.67
124	R	3	26.66	0.38	16.67
124	R	9	26.5	1.03	0
124	Z	1	27.83	1.19	0
124	Z	3	63.06	4.55	3.85
124	Z	9	67.39	3.13	9.52
128	F	0	1.53	0	0
128	G	1	0.65	0	0
128	G	3	1.43	0	0
128	G	9	0.71	0	0
128	R	1	0	0	0
128	R	3	0.32	0	0
128	R	9	0	0	0
128	Z	1	0	0	0
128	Z	3	0.6	0	0
128	Z	9	0.02	0	0
132	F	0	99.96	0.19	50
132	G	1	99.96	0.57	0
132	G	3	99.96	0.13	100
132	G	9	99.96	0.19	50
132	R	1	99.15	0.58	0
132	R	3	99.77	0.44	0
132	R	9	99.73	0.17	0
132	Z	1	99.96	0.06	0
132	Z	3	99.96	0	0
132	Z	9	99.96	0.06	100

Appendix 8 - CheckM results for MD-*Campylobacter* genomes, these values are not standardised to reads in and represent the full sequencing yield of each sample

Stool Id	Condition	Timepoint	Completeness	Contamination	Strain heterogeneity
130r1	F	0	0.38	0	0
130r1	G	1	0.67	0	0
130r1	G	3	0.57	0	0
130r1	G	9	0.02	0	0
130r1	R	1	0	0	0
130r1	R	3	0.48	0	0
130r1	R	9	0.6	0	0
130r1	Z	1	0	0	0
130r1	Z	3	0.76	0	0
130r1	Z	9	0.38	0	0
130r2	F	0	0.08	0	0
130r2	G	1	1.15	0	0
130r2	G	3	0.06	0.06	100
130r2	G	9	0	0	0
130r2	R	1	0.57	0	0
130r2	R	3	0	0	0
130r2	R	9	0.3	0	0
130r2	Z	1	0.02	0	0
130r2	Z	3	0	0	0
130r2	Z	9	0.19	0	0
136	F	0	99.39	0.81	0
136	G	1	83.32	4.03	3.45
136	G	3	80.11	3.9	18.52
136	G	9	82.16	4.3	7.14
136	R	1	47.75	1.54	8.33
136	R	3	50.72	1.65	10
136	R	9	66.63	2.84	12.5
136	Z	1	20.12	0.3	0
136	Z	3	15.7	0.13	0
136	Z	9	8.88	0	0

Appendix 8 - CheckM results for MD-*Campylobacter* genomes, these values are not standardised to reads in and represent the full sequencing yield of each sample

Stool Id	Condition	Timepoint	Completeness	Contamination	Strain heterogeneity
135r1	F	0	99.86	0.25	50
135r1	G	1	99.86	0.13	0
135r1	G	3	99.86	0.25	50
135r1	G	9	99.96	0.13	0
135r1	R	1	97.69	1.92	0
135r1	R	3	99.09	1.14	20
135r1	R	9	99.07	0.49	0
135r1	Z	1	99.29	0.98	12.5
135r1	Z	3	98.02	0.7	40
135r1	Z	9	99.14	1.57	12.5
135r2	F	0	99.86	0.13	0
135r2	G	1	99.86	0.13	0
135r2	G	3	99.96	0.25	0
135r2	G	9	99.86	0.13	0
135r2	R	1	98.02	0.25	0
135r2	R	3	99.04	0.49	0
135r2	R	9	97.11	1.63	37.5
135r2	Z	1	99.9	0.52	0
135r2	Z	3	99.81	1.3	33.33
135r2	Z	9	99.02	0.97	14.29
136	F	0	99.39	0.81	0
136	G	1	83.32	4.03	3.45
136	G	3	80.11	3.9	18.52
136	G	9	82.16	4.3	7.14
136	R	1	47.75	1.54	8.33
136	R	3	50.72	1.65	10
136	R	9	66.63	2.84	12.5
136	Z	1	20.12	0.3	0
136	Z	3	15.7	0.13	0
136	Z	9	8.88	0	0

Appendix 8 - CheckM results for MD-*Campylobacter* genomes, these values are not standardised to reads in and represent the full sequencing yield of each sample

Stool Id	Condition	Timepoint	Completeness	Contamination	Strain heterogeneity
141	F	0	98.5	2.39	11.11
141	G	1	0.19	0	0
141	G	3	8.32	0.06	0
141	G	9	0.13	0	0
141	R	1	88.98	3.9	9.09
141	R	3	0.08	0	0
141	R	9	89.99	3.87	23.81
141	Z	1	2.66	0.13	0
141	Z	3	16.69	0.19	0
141	Z	9	23.95	0.51	0
143	F	0	99.02	0.72	0
143	G	1	99.96	1.06	16.67
143	G	3	99.96	0.14	0
143	G	9	99.96	0.38	0
143	R	1	99.96	0.16	33.33
143	R	3	99.58	1.23	16.67
143	R	9	99.96	0.72	16.67
143	Z	1	3.85	0	0
143	Z	3	4.87	0.1	0
143	Z	9	2.15	0	0
144	F	0	12.19	0.42	50
144	G	1	2.33	0	0
144	G	3	4.1	0	0
144	G	9	1.08	0	0
144	R	1	0.89	0	0
144	R	3	0.33	0	0
144	R	9	0	0	0
144	Z	1	1.23	0	0
144	Z	3	0.1	0	0
144	Z	9	1.49	0	0

Appendix 8 - CheckM results for MD-*Campylobacter* genomes, these values are not standardised to reads in and represent the full sequencing yield of each sample

Stool Id	Condition	Timepoint	Completeness	Contamination	Strain heterogeneity
145	F	0	3.67	0	0
145	G	1	0.38	0	0
145	G	3	0.38	0	0
145	G	9	0	0	0
145	R	1	0	0	0
145	R	3	0	0	0
145	R	9	0	0	0
145	Z	1	3.81	0	0
145	Z	3	3.33	0	0
145	Z	9	4.85	0	0
146	F	0	81.18	4.75	10
146	G	1	73.65	5.93	6.06
146	G	3	59.92	3.73	15
146	G	9	31.26	0.43	33.33
146	R	1	62.18	4.35	8.33
146	R	3	31.6	0.74	0
146	R	9	31.36	1.11	0
146	Z	1	7.84	0	0
146	Z	3	5.22	0	0
146	Z	9	11.49	0	0
147	F	0	98.67	1.77	57.14
147	G	1	64.23	4.61	9.09
147	G	3	75.39	3.79	10.34
147	G	9	16.29	0	0
147	R	1	38.73	1.63	22.22
147	R	3	65.61	6.04	6.9
147	R	9	19.46	0.06	0
147	Z	1	0.99	0	0
147	Z	3	0.59	0	0
147	Z	9	0.08	0	0

Appendix 9 - Wilcoxon rank-sum test results (with Benjamini-Hochberg correction) for CheckM genome completeness of MD-*Campylobacter* genomes in storage conditions R, G, Z and timepoints 1, 3, 9, compared to the pre storage baseline F0.

Comparison	Statistic	Raw p-value	BH-corrected p-value	Significant (BH)
F0 vs G1	0.4594768	0.645891808	0.947844387	FALSE
F0 vs G3	0.32163376	0.747730165	0.947844387	FALSE
F0 vs G9	0.73516288	0.462240302	0.947844387	FALSE
F0 vs R1	0.64326752	0.520050527	0.947844387	FALSE
F0 vs R3	1.65411648	0.098103848	0.947844387	FALSE
F0 vs R9	1.148692	0.250683005	0.947844387	FALSE
F0 vs Z1	1.70006416	0.089118857	0.947844387	FALSE
F0 vs Z3	1.19463968	0.232227838	0.947844387	FALSE
F0 vs Z9	1.33248272	0.182701615	0.947844387	FALSE
G1 vs G3	-0.2297384	0.818295054	0.947844387	FALSE
G1 vs G9	0.55137216	0.581378581	0.947844387	FALSE
G1 vs R1	0.13784304	0.890364468	0.947844387	FALSE
G1 vs R3	1.33248272	0.182701615	0.947844387	FALSE
G1 vs R9	0.78111056	0.434737471	0.947844387	FALSE
G1 vs Z1	0.6892152	0.490687852	0.947844387	FALSE
G1 vs Z3	0.59731984	0.55029386	0.947844387	FALSE
G1 vs Z9	0.59731984	0.55029386	0.947844387	FALSE
G3 vs G9	0.6892152	0.490687852	0.947844387	FALSE
G3 vs R1	0.41352912	0.679218992	0.947844387	FALSE
G3 vs R3	1.24058736	0.214758223	0.947844387	FALSE
G3 vs R9	0.82705824	0.408204051	0.947844387	FALSE
G3 vs Z1	1.24058736	0.214758223	0.947844387	FALSE
G3 vs Z3	0.73516288	0.462240302	0.947844387	FALSE
G3 vs Z9	0.9189536	0.358119842	0.947844387	FALSE
G9 vs R1	-0.4135291	0.679218992	0.947844387	FALSE
G9 vs R3	0.55137216	0.581378581	0.947844387	FALSE
G9 vs R9	0.32163376	0.747730165	0.947844387	FALSE
G9 vs Z1	0.32163376	0.747730165	0.947844387	FALSE
G9 vs Z3	-0.0918954	0.926781178	0.947844387	FALSE
G9 vs Z9	0.04594768	0.963351951	0.963351951	FALSE
R1 vs R3	0.59731984	0.55029386	0.947844387	FALSE
R1 vs R9	0.48245064	0.629485854	0.947844387	FALSE
R1 vs Z1	0.59731984	0.55029386	0.947844387	FALSE
R1 vs Z3	0.62029368	0.535064454	0.947844387	FALSE
R1 vs Z9	0.41352912	0.679218992	0.947844387	FALSE
R3 vs R9	-0.3216338	0.747730165	0.947844387	FALSE
R3 vs Z1	0.2297384	0.818295054	0.947844387	FALSE
R3 vs Z3	-0.2297384	0.818295054	0.947844387	FALSE
R3 vs Z9	-0.2297384	0.818295054	0.947844387	FALSE

Appendix 9 - Wilcoxon rank-sum test results (with Benjamini-Hochberg correction) for CheckM genome completeness of MD-*Campylobacter* genomes in storage conditions R, G, Z and timepoints 1, 3, 9, compared to the pre storage baseline F0.

Comparison	Statistic	Raw p-value	BH-corrected p-value	Significant (BH)
R9 vs Z1	0.50542448	0.613260728	0.947844387	FALSE
R9 vs Z3	0.16081688	0.87223763	0.947844387	FALSE
R9 vs Z9	0.13784304	0.890364468	0.947844387	FALSE
Z1 vs Z3	-0.0918954	0.926781178	0.947844387	FALSE
Z1 vs Z9	-0.3216338	0.747730165	0.947844387	FALSE
Z3 vs Z9	0.13784304	0.890364468	0.947844387	FALSE